Optimising referential coherence in text generation

Rodger Kibble*
University of London

Richard Power†
University of Brighton

This paper describes an implemented system which uses centering theory for planning of coherent texts and choice of referring expressions. We argue that text and sentence planning need to be driven in part by the goal of maintaining referential continuity and thereby facilitating pronoun resolution: obtaining a favourable ordering of clauses, and of arguments within clauses, is likely to increase opportunities for non-ambiguous pronoun use. Centering theory provides the basis for such an integrated approach. Generating coherent texts according to centering theory is treated as a constraint satisfaction problem. We report on two empirical studies: a paired-comparison experiment to test whether readers prefer texts that adhere to centering constraints, and what we believe to be a novel method of corpus analysis involving perturbances, to investigate whether authors aim to promote referential coherence and if so, which centering constraints are more important.

1 Overview

A central task for NLG systems is to produce text which is *coherent*, in a sense in which (1a) is noticeably more coherent than (1b):

1. a. Elixir is a white cream.
   It is used in the treatment of cold sores.
   It contains aliprosan.
   Aliprosan relieves viral skin disorders.
   b. Elixir contains aliprosan.
   Viral skin disorders are relieved by aliprosan.
   Elixir is used in the treatment of cold sores.
   It is a white cream.

We can observe various ways in which text organisation influences coherence: the sequence in which certain facts are presented, the order in which entities are mentioned in a clause, and the possibilities available for identifying the intended reference of pronouns. Generally, (1a) seems to conform better to a reader's expectations of what will be referred to next and of how to resolve underspecified referring expressions, in particular pronouns. These are questions which the well-known Centering Theory (CT) of (Grosz, Joshi, and Weinstein, 1995, henceforth GJW) is concerned with; indeed CT has been highly influential on computational work in anaphora resolution (Brennan, Friedman, and Pollard, 1987), though until recently its potential for improving the output of NLG systems has been relatively unexplored. The starting point of the work reported in this

* Department of Computing, Goldsmiths College, University of London, London SE14 6NW, U.K.
† Information Technology Research Institute, University of Brighton, Brighton BN2 4GJ, U.K.

paper is that NLG systems need some principled basis for making these decisions, and CT has the advantage of providing guidance on the ordering of clauses and arguments in addition to choice of referring expressions. Previous algorithms for pronominalisation such as those of (McCoy and Strube, 1999; Henschel, Cheng, and Poesio, 2000) have addressed the task of deciding whether to realise an entity as a pronoun on the basis of given factors such as its syntactic role and discourse history within a given text structure; what is essentially novel in our approach is that we treat referential coherence as a *planning* problem, on the assumption that obtaining a favourable ordering of clauses, and of arguments within clauses, is likely to increase opportunities for non-ambiguous pronoun use. Centering theory provides the basis for such an integrated approach[1].

This paper describes a method for applying CT to problems in text planning and pronominalisation in order to improve the fluency and readability of generated texts. This approach is applicable in principle to any system which produces hierarchically structured text plans using a theory of coherence relations, with the following additional assumptions:

- "shallow" lexicalisation with effectively a one-to-one correspondence between predicates and verbs, so that the options for syntactic realisation can be predicted from the argument structure of predicates. This appears to be standard in applied NLG systems (Cahill, 1999).
- pronominalisation is deferred until grammatical relations and word order have been determined.

Our exposition will refer to an implemented document generation system, ICONOCLAST, which uses the technique of *constraint satisfaction* (van Hentenryck, 1989; Power, 2000; Power, Scott, and Bouayad-Agha, 2003) with CT principles implemented among a set of soft constraints. The ICONOCLAST system allows the user to specify content and rhetorical structure through an interactive knowledge-base editor, and supports fine-grained control over stylistic and layout features. The user-determined rhetorical structure is transformed into a *text structure* or a set of candidate text structures which respect various text-formation rules encoded as hard constraints. Not all of the resulting text structures will give rise to stylistically acceptable documents, and of those which may be judged acceptable some will be noticeably preferable to others. The text structuring phase is followed by an evaluation of the candidate structures, where they are ranked according to a set of preferences encoded as soft constraints. Centering preferences are weighted along with other stylistic constraints to fix the preferred final ordering both of propositions in the text and of arguments within a clause. So for example Figure 1 is a plausible rhetorical structure for both (1a) and (1b); one difference between them is that (1a) has fewer violations of centering preferences.

It is not our primary aim in this short paper to provide an empirical assessment of the claims of CT, for which we refer the reader to the relevant papers such as those collected in (Walker, Joshi, and Prince, 1998) as well as (Poesio et al., 2002) and other works cited there. However, we recognise the need to address some specific empirical issues: can it be shown that CT constraints actually make a difference to the acceptability of texts, and how can weights for these constraints be obtained from data? In section 5 we report on ongoing work dealing with these questions. A paired-comparison study of judgments by naïve subjects indicates that the first question can be answered in the affirmative, and a corpus study using what we believe to be a novel technique involving

1 Callaway and Lester (2002) complain that CT-based pronominalisation algorithms "assume that the discourse tree was constructed with Centering theory in mind"; in our case this assumption is justified.

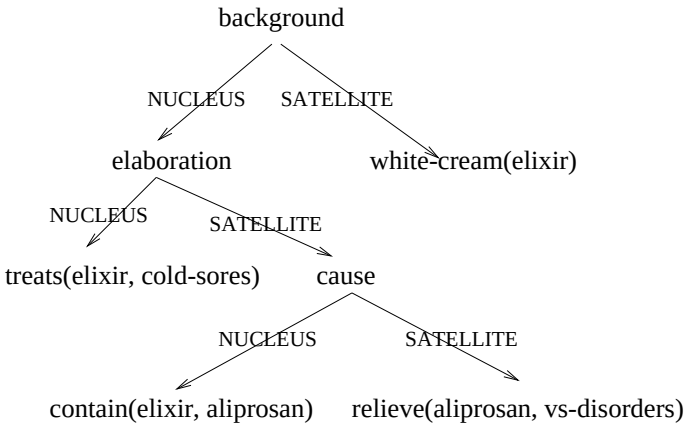

Figure 1
Rhetorical structure for example 1.

perturbations provides clear evidence of preferences between the different constraints. Results to date suggest that further empirical research will be worthwhile; in fact one of the strengths of our framework is that it can be used as a research tool for the evaluation of variants of CT, as different realisations of an input sequence can be generated by varying control parameters and one can very quickly see the results of alternative choices.

1.1 Related Work

Other researchers have applied CT to generation, though to our knowledge none have applied it to text planning, sentence planning and pronominalisation in the integrated way that we present in this paper. This general approach is anticipated by McKeown's (1985) approach to text planning, where referential coherence is taken to be one of the factors determining fluency, though her system predates RST and Centering. REFS: MELLISH, MARCU... Mittal et al. (1998) apply what we term **salience** to sentence planning, with the goal of realising the Cb as Subject, though the text planner does not have a goal of attempting to maintain the same Cb. We regard Cheng's (2000) work on the interaction of centering preferences and aggregation in text planning as complementary to our enterprise. (Karamanis, 2001; Kibble, 2001; Beaver, 2003) have argued for a *ranking* of the centering principles as opposed to weighting, and indeed Beaver (op cit.) provides a unified formulation of the centering rules and constraints as a ranked set of OT constraints. However as will be argued below, we believe that such a ranking stands in need of empirical justification, and the data presented in (Beaver, 2003) actually provides little evidence for ranking as opposed to weighting of constraints (see (Kibble, 2003)). Karamanis and Manurung (2002) report on a series of computational experiments aiming to see how much can be achieved using the principle of Continuity alone, in a particular domain of museum texts consisting largely of extended descriptions of artefacts. Krahmer and Theune's (2002) Modified Incremental Algorithm (MIA) is also relevant; they demonstrate that salience weights can be used to determine whether reduced forms of definite descriptions can be generated, giving CT as one way to determine the weights. The MIA has not been incorporated into the system described here but it provides an obvious extension.

2 Centering Parameters

We assume some familiarity with the basic concepts of CT. In this section we briefly and informally summarise the main assumptions of the theory and explain how we have interpreted and applied these assumptions.

1. For each utterance in a discourse there is said to be at most one entity which is the centre of attention or *center* (**Constraint 1**). The center in an utterance U_n is the most highly ranked entity realised in U_{n-1} which is also realised in U_n (**Constraint 3**). This is also referred to as the *backward-looking center* or *Cb*. (The set of entities mentioned in an utterance U_n is defined by **Constraint 2** as the set of *forward-looking centers* or $Cfs(U_n)$.) It is not entirely clear whether Constraint 1 is to be taken as an empirical claim or as a *stipulation* that some entity must be designated as Cb, if necessary by constructing an indirect anaphoric link.

2. There is a preference for consecutive utterances within a discourse segment to keep the same entity as the center, and for the center to be realised as the highest-ranked entity or *preferred center* (*Cp*). Kibble (1999) dubbed these principles **cohesion** and **salience** respectively. Combinations of these preferences provide the familiar canonical set of *transitions* shown in Figure 2, ranked in the stipulated order of preference first set out as **Rule 2** by (Brennan, Friedman, and Pollard, 1987) and adopted by (Walker, Joshi, and Prince, 1998).

3. The center is the entity which is most likely to be pronominalised: GJW's **Rule 1** in its weakest form states that if any entity is referred to by a pronoun, the *Cb* must be.

CONTINUE: **cohesion** and **salience** both hold; same center (**or** $Cb(U_n)$ undefined), realised as *Cp* in U_{n+1};
RETAIN: **cohesion** only; i.e. center remains the same but is not realised as *Cp* in U_{n+1};
SMOOTH SHIFT: **salience** only; center of U_{n+1} realised as *Cp* but not equal to $Cb(U_n)$;
ROUGH SHIFT: neither **cohesion** nor **salience** holds.

Figure 2
Centering Transitions

As (Poesio et al., 2002) point out, CT can be viewed as a "parametric" theory in that key notions such as *utterance* and *previous utterance*, *realisation* of entities and *ranking* are not given precise definitions by GJW, and subsequent applied studies have had to begin by fixing particular instantiations of these notions.

Ranking Since (Brennan, Friedman, and Pollard, 1987) a ranking in terms of grammatical roles (or *obliqueness*) has become standard, for example

SUBJECT > DIRECT OBJECT > INDIRECT OBJECT > OTHERS.

We have simplified matters somewhat for the purposes of this implementation. First, we assume that syntactic realization serves only to distinguish the Cp from all other referents, which are ranked on the same level: thus effectively SUBJECT > OTHERS. Secondly, we assume that the system already knows, from the argument structure of the proposition, which entities can occur in subject position: thus in realising a proposition $ban(fda, elixir)$, both arguments are potential Cps because active and passive realizations are both allowed; for $contain(elixir, gestodene)$, only $elixir$ is a potential Cp

because we disallow 'Gestodene is contained by Elixir'.

Realisation GJW's original formulation distinguished between "direct" realisation or coreference, and "indirect" realisation which corresponds to *bridging reference*. As an example, in (1a) the terms "cold sores" and "viral skin disorders" are not strictly coreferential and so do not count as direct realisations of the same entity, but if we allow indirect realisation then there is the potential for one of these to be identified as *Cb*, in a sequence such as . . . *Elixir is used to treat cold sores. Viral skin disorders are relieved by aliprosan*. Again, we keep things simple at this stage by only treating nominal expressions as realisations of the same entity if they strictly co-refer. As (Poesio et al., 2002) observe, under this interpretation of realisation a number of utterances will lack an identifiable Cb, so we have to allow for a "no-Cb" transition in addition to the canonical transitions listed in Figure 2.

Utterance and previous utterance Two different approaches to the realisation of "utterance" have become associated with the work of Kameyama (1998) and Suri et al (1999). To simplify somewhat, Kameyama argued that the local focus is updated in a linear manner by tensed clauses rather than by sentences, while Suri et al present evidence that the subject of the main clause in a complex sentence is likely to be the preferred antecedent for a subject pronoun in an immediately following sentence, winning out over candidates in an intervening subordinate clause as in example 2:

2. Dodge^i was robbed by an ex-convict^j the other night.

   The ex-convict$_j$ tied him$_i$ up because he$_i$ wasn't cooperating.

   Then he$_j$ took all the money and ran / #he$_i$ started screaming for help.

In fact we would argue that Suri et al's analysis does not establish whether the accessibility effects are due to the syntactic or the rhetorical structure of utterances. The examples they present all involve sentences of the form *Sx because Sy* corresponding to the rhetorical pattern *Nucleus - connective - Satellite*. Their results are therefore consistent with the hypothesis that the nucleus of a preceding segment is more accessible than the satellite. We allow the user of our system to choose between two strategies: a linear, Kameyama-style approach or a hierarchical approach where the utterance is effectively identified with a rhetorical span. Our approach is more general than that of Suri et al. as it covers cases where the components of a complex rhetorical span are realised in different sentences. Veins Theory (Cristea, Ide, and Romary, 1998) provides a possible formalization of the intuition that some earlier propositions become inaccessible as a rhetorical boundary is crossed. The theory could be applied to centering in various ways; we have implemented perhaps the simplest approach, in which centering transitions are assessed in relation to the nearest *accessible* predecessor. In many cases the linear and hierarchical definitions give the same result, but sometimes they diverge, as in the following schematic example:

3. $ban(fda, elixir)$ since $contain(elixir, gestodene)$.

   However, $approve(fda, elixirplus)$.

Following Veins Theory, the predecessor of $approve(fda, elixirplus)$ is $ban(fda, elixir)$; its linear predecessor $contain(elixir, gestodene)$ (an embedded satellite) is inaccessible. This makes a considerable difference: under a hierarchical approach, fda can be the Cb of the final proposition; under a linear approach, this proposition has no Cb.

Transitions vs Constraints Kibble (1999; 2001) argued for a decomposition of the canonical transition types into the principles of **cohesion** and **salience**, partly on the architectural grounds that this made it easier to apply CT to the generation task, and partly on empirical grounds that the preference ordering assumed by GJW was not strongly supported by corpus evidence, and that the transitions were better seen as *epiphenomenal*, emerging in a partial ordering from the interaction of more fundamental constraints. We follow this general approacg, including among the constraints the principle of **continuity**: each utterance should have at least one referent in common with the preceding utterance, which is effectively a restatement of GJW's Constraint 1. If we assign a weight of 1 each to **cohesion** and **salience** and 2 to **continuity** we obtain a partial ordering over the canonical transitions as follows:

$$0 : \textsc{Continue} > 1 : \{\textsc{Retain} \mid \textsc{Smooth Shift}\} > 2 : \{\textsc{Rough Shift} \mid \textsc{No Cb}\}$$

Any relative weighting or ranking of **coherence** over **salience** would need to be motivated by evidence that RETAIN is preferred over SMOOTH SHIFT, and we are not aware of any conclusive evidence in the literature (see (Kibble, 1999) for further discussion).

This approach also means that Strube and Hahn's (1999) principle of **cheapness** can be naturally incorporated as an additional constraint: this is a requirement that $Cp(U_{n-1}) = Cb(U_n)$. The principle of cheapness effectively cashes out the informal definition of the Cp as "represent[ing] a prediction about the Cb of the following utterance" (Walker, Joshi, and Prince, 1998, Ch. 1:3) In classic variants of the theory this only happens indirectly as a result of transition preferences, and only following a Continue or Smooth Shift, since the Cp is also the Cb and Rule 2 predicts that the preferred transition will maintain the same Cb. However the prediction is not entailed by the theory following a Retain, Rough Shift or no-Cb transition, or indeed for the first sentence in a discourse, when there is effectively no prediction concerning the Cp. Strube and Hahn claim that the cheapness principle is motivated by the existence of Retain-Shift patterns which are evidently a common means of introducing a new topic (cf also (Brennan, Friedman, and Pollard, 1987, henceforth BFP)). To summarise, our system incorporates the following constraints:

cohesion: $Cb(U_{n-1}) = Cb(U_n)$

salience: $Cp(U_n) = Cb(U_n)$

cheapness: $Cp(U_{n-1}) = Cb(U_n)$

continuity: $Cfs(U_{n-1}) \cap Cfs(U_n) \neq \emptyset$

Preferences: transitions, pairs or sequences? The original version of GJW's Rule 2 specified that *sequences* of Continue transitions are preferred over sequences of Retains, and so on; in BFP's implementation however transitions are evaluated incrementally and the preference applies to individual transitions such as Continue vs Retain rather than to sequences. Strube and Hahn take an intermediate position: in their formulation, *pairs* of transitions $\langle\langle U_i,\ U_j\rangle,\ \langle U_j,\ U_k\rangle\rangle$ are preferred which are *cheap*, i.e. $Cp(U_j) = Cb(U_k)$. Strube and Hahn intended the preference for cheap transition pairs to replace GJW's Rule 2 in toto, which seems a rather weak requirement. On the other hand the original GJW formulation is difficult to verify since as Poesio et al. (2002, page 66) found, sequences of multiple occurrences of the same transition type turn out to be relatively rare. Our position is a little more complex as we do not directly aim to generate particular transitions or sequences of transitions but to minimise violations of the constraints **continuity**, **cohesion**, **salience** and **cheapness**. Violations are computed on individual

nodes and summed for each candidate text structure, so we may expect that the candidate with the fewest violations will have a preponderance of the preferred transitions. The system is certainly more slanted towards global optimisation than BFP's incremental model, but may be said to achieve this in a more natural way than a strategy of trying to produce uniform sequences of transitions.

Pronominalisation GJW's Rule 1 is rather weak as a guide to pronominalisation decisions in general, as it only mentions the Cb and gives little guidance on when or whether to pronominalise non-Cbs. An important consideration for NLG is to minimise the possibility of ambiguity and so we adopt a cautious strategy: the user can choose between invariably pronominalising the Cb or using a fairly simple algorithm based on parallelism of grammatical roles. A possible future development is to supplement our CT-based text planner with a more sophisticated pronominalisation algorithm such as that of (Henschel, Cheng, and Poesio, 2000).

3 Generation issues

CT has developed primarily in the context of natural language interpretation, focussing on anaphora resolution (see e.g., (Brennan, Friedman, and Pollard, 1987)). As stated above, the novel contribution of this paper is an integrated treatment of pronominalisation and *planning*, aiming to determine whether the principles underlying the constraints and rules of the theory can be "turned round" and used as planning operators for generating coherent text. We have assumed some familiarity in the foregoing with terms such as "text planning" and "sentence planning". These are among the distinct tasks identified in Reiter's "consensus architecture" for Natural Language Generation (Reiter, 1994):

Text Planning/Content Determination - deciding the content of a message, and organising the component propositions into a text structure (typically a tree).

Sentence Planning - aggregating propositions into clausal units and choosing lexical items corresponding to concepts in the knowledge base; this is the level at which the order of arguments and choice of referring expressions will be determined.

Linguistic realisation - surface details such as agreement, orthography etc.

Reiter observed that these functions can often be identified with discrete modules in applied NLG systems, and that a de facto standard had emerged where these modules are organised in a *pipeline* such that information flows in one direction, and only between modules which immediately succeed each other.

Breaking down the generation task in this way makes it evident that there are various ways the distinct principles of CT can be incorporated. **Continuity** and **cohesion** naturally come under Text Planning: respectively, ordering a sequence of utterances to ensure that each has a backward-looking center, and maintaining the same entity as the center within constraints on ordering determined by discourse relations. **Salience** and **cheapness** on the other hand would come under Sentence Planning since in each case a particular entity is to be realised as Subject. However, we encounter an apparent paradox in that identifying the center itself depends on grammatical salience as determined by the Sentence Planner - for example, choice of active or passive voice. Consequently, the text planner appears to rely on decisions made at sentence planning level, which is incompatible with the fact that

> pipelined systems cannot perform general search over a decision space which includes decisions made in more than one module.
>
> (Reiter, 2000, page 000)

We can envisage three possibilities for incorporating CT into a generation architecture:

1. "Incremental" sentence-by-sentence generation, where the syntactic structure of U_n is determined before the semantic content of U_{n+1} is planned. That is, the Text Planner would plan the content of U_{n+1} by aiming to realise a proposition in the knowledge base which mentions an entity which is salient in U_n. We are not aware of any system which performs all stages of generation in a sentence-by-sentence way, and in any case this type of architecture would not allow the cheapness principle to be implemented as it would not support the necessary forward planning. This would also limit the possibilities for evaluation of candidate outputs, as the system would not be able to evaluate a multi-sentence sequence.

2. A pipelined system where the "topic" or "theme" of a sentence is designated independently as part of the semantic input, and centering rules reflect the information structure of a discourse. This approach was suggested by (Kibble, 1999), who proposed that text and sentence planning should be driven by the goal of realising the designated topic in positions where it will be interpreted as the Cb. However, this is not really a solution so much as a refinement of the problem, since it simply shifts the problem of identifying the topic. (Prince, 1999) notes that definitions of "topic" in the literature do not provide objective tests for topichood and proposes that the topic should be identified with the centre of attention as defined by CT; however, what would be needed here would be a more fundamental definition which would account for a particular entity being chosen to be the centre of attention in the first place.

3. The solution we adopt is to treat the task of identifying Cbs and Cps as an optimisation problem. We assume that certain options for syntactic realisation can be predicted on the basis of the argument structure of predicates, which means that centering constructs can be calculated as part of Text Planning *before* syntactic realisation takes place, and the paradox noted above is resolved. Pronominalisation decisions are deferred until a point where grammatical relations and word order have been fixed.

4 Generation as Constraint Satisfaction

In this section we give an overview of our text planning component in order to set the implementation of CT in context. The methodology is more fully described in (Power, Scott, and Bouayad-Agha, 2003).

The text planner was developed within ICONOCLAST, a project which investigated applications of constraint-based reasoning in Natural Language Generation using as subject-matter the domain of medical information leaflets. Following (Scott and de Souza, 1990), we represent rhetorical structure by graphs like figure 3, in which non-terminal nodes represent RST relations, terminal nodes represent propositions, and linear order is unspecified. The task of the text planner is to realize the rhetorical structure as a *text structure* in which propositions are ordered, assigned to textual units (e.g., sentences, paragraphs, vertical lists), and linked where appropriate by discourse connectives (e.g., 'since', 'however'). The boundary between text and sentence planning is drawn at the realisation of elementary propositions rather than the generation of individual sentences. If a rhetorical subtree is realised as a complex sentence, the effect is that "text planning" trespasses into the higher-level syntax of the sentence, leaving only

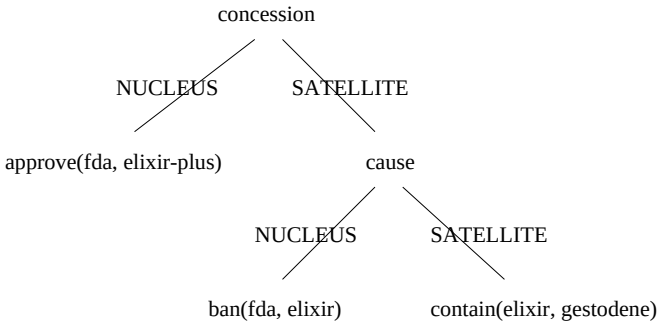

Figure 3
Rhetorical structure

the elementary propositions to be realised by "sentence planning"[2].

Even for a simple rhetorical input like figure 3 many reasonable text structures can be generated. Since there are two nucleus-satellite relations, the elementary propositions can be ordered in four ways; several discourse connectives can be employed to realize each rhetorical relation (e.g. *concession* can be realized by 'although', 'but' and 'however'); at one extreme, the text can be spread out over several paragraphs, while at the other extreme it can be squeezed into a single sentence. With fairly restrictive constraint settings, the system generates 24 text-structure patterns for figure 3, including the following (shown schematically):

A. Since $contain(elixir, gestodene)$, $ban(fda, elixir)$.
However, $approve(fda, elixirplus)$.

B. $approve(fda, elixirplus)$, although since $contain(elixir, gestodene)$, $ban(fda, elixir)$.

The final output texts will depend on how the propositions are realized syntactically; among other things this will depend on centering choices within each proposition.

In outline, the procedure that we propose is as follows:

1. Enumerate all text structures that are acceptable realizations of the rhetorical structure.
2. For each text structure, enumerate all permissible choices for the Cb and Cp of each proposition.
3. Evaluate the solutions, taking account of referential coherence among other considerations, and choose the best.

For the example in figure 3, centers can be assigned in four ways for each text-structure pattern, making a total of 96 solutions.

As will probably be obvious, such a procedure could not be applied for rhetorical structures with many propositions. For examples of this kind, based on the relations 'cause' and 'concession' (each of which can be marked by several different connectives),

2 See (Power, Scott, and Bouayad-Agha, 2003) for detailed motivation of this concept of text structure as as a level of representation distinct from both rhetorical structure and syntactic structure.

we find that the total number of text-structures is approximately 5^{N-1} for N propositions. Hence with $N = 5$ we would expect around 600 text structures; with perhaps 5-10 ways of assigning centers to each text structure, the total number of solutions would approximate to 5000. Global optimization of the solution therefore becomes impracticable for texts longer than about five propositions; we address this problem by a technique of *partial optimization* in which a high-level planner fixes the large-scale structure of the text, thus defining a set of local planning problems each small enough to be tackled by the methods described here.

Stage 1 of the planning procedure is described in more detail in (Power, Scott, and Bouayad-Agha, 2003). A brief summary follows, after which we focus on stages 2 and 3, in which the text planner enumerates the possible assignments of centers and evaluates which is the best.

4.1 Generating and evaluating text structures

A text structure is defined in ICONOCLAST as an ordered tree in which each node has a feature named TEXT–LEVEL. Values of TEXT–LEVEL are represented by integers in the range $0 \ldots L_{Max}$; these may be interpreted in various ways, but we will assume here that $L_{Max} = 4$ and that integers are paired with descriptive labels as follows:

0 text-phrase

1 text-clause

2 text-sentence

3 paragraph

4 section

Informally, a text structure is well-formed if it respects the hierarchy of textual levels, so that sections are composed of paragraphs, paragraphs of text–sentences, and so forth. An example of an ill-formed structure would be one in which a text-sentence contained a paragraph; such a structure can occur only when the paragraph is indented — a possibility we are excluding here. As well as being a well-formed text structure, a candidate solution must realize a rhetorical structure 'correctly', in a sense that we need to make precise. Roughly, a correct solution should satisfy three conditions:

1. The terminal nodes of the TS should express all the elementary propositions in the RS; they may also contain discourse connectives expressing rhetorical relations in the RS, although for some relations discourse connectives are optional.

2. The TS must respect rules of syntax when it combines propositions and discourse connectives within a text-clause; for instance, a conjunction such as 'but' linking two text-phrases must be coordinated with the second one.

3. The TS must be structurally compatible with the RS.

The first two conditions are straightforward, but what is meant by 'structural compatibility'? We suggest the crucial criterion should be as follows: **any grouping of the elementary propositions in the TS must also occur in the RS**. In other words, the text-structurer is allowed to eliminate groupings, but not to add any. More formally:

- If a node in the TS dominates terminal nodes expressing a set of elementary propositions, there must be a corresponding node in the RS dominating the same set of propositions.
- The converse does not hold: for instance, an RS of the form $R_1(R_2(p_1, p_2), p_3)$ can be realized by a paragraph of three sentences, one for each proposition, even though this TS contains no node dominating the propositions $(p_1 and p_2)$ that are grouped by R_2. However, when this happens, the propositions grouped together in the RS must remain consecutive in the TS; solutions in which p_3 comes in between p_1 and p_2 are prohibited.

Our procedure for generating candidate solutions is based on a technique for formulating text structuring as a constraint satisfaction problem (CSP) (Hentenryck, 1989), using the ECLIPSE logic programming environment[3]. In general, a CSP is characterized by the following elements:

- A set of variables $V_1 \ldots V_N$.
- For each variable V_i, a finite domain D_i of possible values.
- A set of constraints on the values of the variables. (For integer domains these often use 'greater than' and 'less than'; other domains usually rely on 'equal' or 'unequal'.)

A solution assigns to each variable V_i a value from its domain D_i while respecting all constraints. For instance each node of rhetorical structure is annotated with a TEXT-LEVEL variable with the domain $0 \ldots L_{Max}$ and an ORDER variable with the domain $1 \ldots N$ where N is the number of sisters. Some applicable constraints are listed below. Depending on the constraints, there may be multiple solutions, or there may be no solution at all. We distinguish between *hard constraints* which are applied during the generation phase, determining which candidate structures can be generated, and *soft constraints* which apply during an evaluation phase in which the enumerated solutions are ordered from best to worst. Some examples of hard and soft constraints are shown in Table 1.

As an example of the **Rhetorical Grouping** constraint: in the following sequence, the tree structure shown in Figure 3 is not reflected in the syntactic structure and so other things being equal, this will be less preferred than any of the variants shown in Table 2.

(a) The FDA bans Elixir. **(b)** It contains gestodene. **(c)** However the FDA approves Elixir–plus.

4.2 Choosing centers

Given a text structure, we enumerate all permissible centering assignments as follows:

1. Determine the predecessor U_{n-1} (if any) of each proposition U_n.
2. List the potential Cbs and Cps of each proposition, henceforth denoted by ΣCb and ΣCp.
3. Compute all combinations from ΣCb and ΣCp that respect the fundamental centering constraint that $Cb(U_n)$ should be the most salient candidate in U_{n-1}.

3 See `http://www-icparc.doc.ic.ac.uk/eclipse/`

Name	Type	Description
Root Domination $L_p > L_d$	Hard	The TEXTLEVEL of the root node r must exceed that of any daughter d.
Parental Domination $L_p \geq L_d$	Hard	The TEXTLEVEL of a parent node p must be equal to or greater than the textlevel of any daughter d.
Sister Equality $L_a = L_b$	Hard	If nodes a and b are descended from the same parent, they must have the same TEXTLEVEL.
Sister Order $O_a \neq O_b$	Hard	If nodes a and b are descended from the same parent, they must have different values of ORDER.
Rhetorical grouping	Soft	Failure to express a rhetorical grouping can be treated as a defect.
Oversimple paragraph	Soft	A paragraph containing only one text-sentence can be treated as a defect.
Centering	Soft	Constraints derived from Centering Theory.

Table 1
Examples of text structuring constraints

As stated earlier, two criteria for determining the predecessor have been implemented; the user can select one or other criterion, thus using the NLG system to test different approaches. Following a *linear* criterion, the predecessor is simply the proposition that precedes the current proposition in the text, regardless of structural considerations. Following a *hierarchical* criterion, the predecessor is the most accessible previous proposition, in the sense defined by Veins Theory (Cristea et al 1998). For now we assume the criterion is linear.

$\Sigma Cb(U_n)$ (potential Cbs of proposition U_n) is given by the intersection between $Cf(U_n)$ and $Cf(U_{n-1})$ — i.e., all the referents they have in common. The potential Cps are those referents in the current proposition that can be realized as most salient. Obviously this should depend on the linguistic resources available to the generator; the system actually uses a simpler rule based on argument types within the proposition. Figure 4 shows the potential Cbs and Cps for the proposition sequence in solution A. As

U	**Proposition**	$\Sigma Cb(U)$	$\Sigma Cp(U)$
U_1	contain(elixir, gestodene)	[ ]	[elixir]
U_2	ban(fda, elixir)	[elixir]	[fda, elixir]
U_3	approve(fda, elixir-plus)	[fda]	[fda, elixir-plus]

Figure 4
Cbs and Cps for solution A.

stated earlier, our treatment of salience here simplifies in two ways: we assume that syntactic realization serves only to distinguish the Cp from all other referents, and that the system already knows, from the argument structure of the proposition, which entities can occur in subject position. With these simplifications, the enumeration of centering assignments is straightforward; in the above example, four combinations are possible, since there are two choices each for $Cp(U_2)$ and $Cp(U_3)$.

4.3 Evaluating solutions

The system evaluates candidate solutions by applying a battery of tests to each node of the text plan. Each test identifies whether the node suffers from a particular defect.

For instance, one stylistic defect (at least for the rhetorical relations occurring in figure 3) is that of placing nucleus before satellite; in general, the text reads better if important material is placed at the end. For each type of defect, we specify a weight indicating its importance: in evaluating continuity of reference, for example, the defect 'No Cb' is regarded as more significant than other defects. Other violations are only recorded in the case where a Cb is present, so if all violations were weighted equally this could result in a 'no-Cb' transition being treated as less serious than an "expensive" Smooth Shift, for example (violating cheapness and cohesion). Summing the weighted costs for all defects, we obtain a total cost for the solution; our aim is to find the solution with the lowest total cost.

Regarding centering, the tests currently applied are as follows.

Salience violation
A proposition U_n violates salience if $Cb(U_n) \neq Cp(U_n)$. This defect is assessed only on propositions that have a backward-looking center.

Cohesion violation
A transition $\langle U_{n-1},\ U_n\rangle$ violates cohesion if $Cb(U_n) \neq Cb(U_{n-1})$. This defect is not recorded when either U_n or U_{n-1} has no Cb.

Cheapness violation
A transition $\langle U_{n-1},\ U_n\rangle$ violates cheapness if $Cb(U_n) \neq Cp(U_{n-1})$. This defect is assessed only on propositions that have a backward-looking center.

Continuity violation
This defect is recorded for any proposition with no Cb, except the first proposition in the sequence (which by definition cannot have a Cb).

Relative weightings for these defects can be chosen by the user; for the current examples we have chosen a neutral scheme with a weight of 3 for continuity violations and 1 each for the others, so that a no-Cb transition is ranked equally bad as an "expensive" Rough Shift. Applied to the four solutions to text structures A and B these definitions yield costs shown in Table 2. According to our metric, solutions A1 and A2 should be preferred because they incur less cost than any others, with B3 and B4 the least preferred.

Although this paper focusses on centering issues, it is important to remember that other aspects of text quality are evaluated at the same time: the aim is to compute a global measure so that disadvantages in one factor can be weighed against advantages in another. For instance, text pattern B is bound to yield poor continuity of reference because it orders the propositions so that U_1 and U_2 have no referents in common. Text pattern A avoids this defect, but this does not automatically mean that A is better than B; there may be other reasons, unconnected with centering, for preferring B to A. The constraints which have an effect on clause ordering include:

Satellite before nucleus For nucleus-satellite relations, place the satellite before the nucleus.

Right-branching structure If an elementary proposition is coordinated with a complex rhetorical structure, place the elementary proposition first.

Centering constraints penalise orderings which violate centering preferences

Text pattern B is favoured by **Right-branching structure**, but in this case the centering constraints will "conspire" with **Satellite before nucleus** to favour pattern A overall.

Solution	Text	Cb	Cp	**Defects**	**Total**
A1	Since Elixir contains gestodene the FDA bans Elixir. However, it approves Elixir+.	∅ elixir fda	elixir fda fda	none sal coh	2
A2	Since Elixir contains gestodene it is banned by the FDA. However, the FDA approves Elixir+.	∅ elixir fda	elixir elixir fda	none none coh, ch	2
A3	Since Elixir contains gestodene the FDA bans Elixir. However, Elixir+ is approved by the FDA.	∅ elixir fda	elixir fda elixir+	none sal sal, coh	3
A4	Since Elixir contains gestodene it is banned by the FDA. However, Elixir+ is approved by the FDA.	∅ elixir fda	elixir elixir elixir+	none none sal, coh, ch	3
B1	The FDA approves Elixir+ although since Elixir contains gestodene it is banned by the FDA.	∅ ∅ elixir	fda elixir elixir	none cont none	3
B2	Elixir+ is approved by the FDA although since Elixir contains gestodene it is banned by the FDA.	∅ ∅ elixir	elixir+ elixir elixir	none cont none	3
B3	The FDA approves Elixir+ although since Elixir contains gestodene the FDA bans Elixir.	∅ ∅ elixir	fda elixir fda	none cont sal	4
B4	Elixir+ is approved by the FDA although since Elixir contains gestodene the FDA bans Elixir.	∅ ∅ fda	elixir+ elixir elixir	none cont sal	4

Table 2
Realisations of text patterns A and B, with weights: cohesion | salience | cheapness = 1, continuity = 3

5 Evaluation

We carried out two empirical studies to determine whether centering constraints really do make a difference to the acceptability of texts and if so, whether some constraints have more effect than others.

5.1 Paired comparison study

Systematically exploring options Are readers sensitive to cohesion, salience, and cheapness, and if so, which violations matter most? Since these constraints are not independent (if two are satisfied in a transition, the third will be as well) and assuming contin-

uous transitions, there are five possible cases:

1. $Cp(U_i) = Cb(U_i) = Cb(U_j)$: All three properties are satisfied; no violations.
2. $Cp(U_i) = Cb(U_i) \neq Cb(U_j)$: Salience satisfied; cohesion and cheapness violated.
3. $Cp(U_i) = Cb(U_j) \neq Cb(U_i)$: Cheapness satisfied; cohesion and salience violated.
4. $Cb(U_i) = Cb(U_j) \neq Cp(U_i)$: Cohesion satisfied; salience and cheapness violated.
5. $Cp(U_i) \neq Cb(U_i) \neq Cb(U_j)$: No properties satisfied; all three violated.

These patterns can be obtained using a three-utterance discourse and at least three referents. Any other referents should occur only once. The second utterance is U_i and the third utterance is U_j.

1. $Cp(U_i) = Cb(U_i) = Cb(U_j)$: Discourse pattern $ab - abc - a$.
2. $Cp(U_i) = Cb(U_i) \neq Cb(U_j)$: Discourse pattern $ba - bac - a$.
3. $Cp(U_i) = Cb(U_j) \neq Cb(U_i)$: Discourse pattern $ba - abc - a$.
4. $Cb(U_i) = Cb(U_j) \neq Cp(U_i)$: Discourse pattern $ab - cab - a$.
5. $Cp(U_i) \neq Cb(U_i) \neq Cb(U_j)$: Discourse pattern $ba - cab - a$.

Examples of the patterns A pilot study has shown that people give a lower rating to solutions that include passive constructions. We therefore prefer to manipulate the ranking of referents within an utterance without using the passive, as in the following examples:

1. Mozart married Constanze in 1782. He lived with Constanza and a maid in Vienna. Later that year he wrote his first piano concerto.
2. Constanze married Mozart in 1782. She lived with him and a maid in Vienna. Later that year he wrote his first piano concerto.
3. Constanze married Mozart in 1782. He lived with her and a maid in Vienna. Later that year he wrote his first piano concerto.
4. Mozart married Constanze in 1782. A maid lived with Mozart and Constanze in Vienna. Later that year he wrote his first piano concerto.
5. Constanze married Mozart in 1782. A maid lived with Mozart and Constanze in Vienna. Later that year he wrote his first piano concerto.

Experimental method Rather than asking subjects to rank all five variants in order, we conducted a Pair Comparison experiment where subjects were presented with all 10 pairs of the five patterns and asked to indicate their preference in each case. Pair Comparison is a popular technique in market research and psychometrics but has not so far been widely used for NLG evaluation[4]. Subjects were also allowed to say both alternatives were equally acceptable, but it turned out that the number of ties was rather small (12 out of 280 judgments, including 3 from one respondent). In order to guard against fatigue or learning effects, and any bias resulting from the order in which the variants were read, a unique questionnaire was generated for each subject. The order in which the pairs were presented, and the order of elements within pairs were randomised, and content words in the examples were varied systematically, giving examples such as

4 We are aware of an evaluation of the STOP system by Reiter et al. (2000) and there have also been evaluations of MT and TTS systems, e.g. (Alvarez and Huckvale, 2002).

```
Passage 1

(a)

 Lauren Bacall first appeared with Humphrey Bogart in To Have And
 Have Not.
 He acted with her and Martha Vickers in The Big Sleep.
 He is best remembered for his role in Casablanca.

(b)

 Lauren Bacall first appeared with Humphrey Bogart in To Have And
 Have Not.
 She acted with him and Martha Vickers in The Big Sleep.
 He is best remembered for his role in Casablanca.

Preferred version: [ ] (Please put "a", "b", or "-" if you have no
preference.)
```

Subjects for this study were all IT professionals who responded to a request posted on the email list of the British Computer Society, Sussex branch. Subjects were thus self-selected; all attested that they were native speakers of English and had not studied linguistics. 28 subjects completed a questionnaire for this experiment which was conducted entirely by email.

Results Initial results were analysed by simply summing the number of times each variant was preferred over another, and performing a sign test on versions which are adjacent in the resulting ranking to check whether the results are significant. A further test would be to repeat the experiment with the same subjects and the same materials but distributed in a different random order, and perform a related-samples sign test on the results (cf. (Alvarez and Huckvale, 2002)). Our initial hypothesis was that Version 1 with no violations would be clearly preferred over all other candidates and that Version 5 with all constraints violated would be the least preferred. The raw data supports the initial hypothesis and suggests a ranking Cheapness > Salience > Cohesion (Table 3). However, sign tests on the direct comparisons show significant preferences for V1 over V3, V3 over V4 and V4 over V5 but not for V3 over V2 or for V2 over V4. (Note that the values for N differ as ties were discounted.)

Version	Constraints satisfied	Preferred
1.	All	99
2.	Salience	51.5
3.	Cheapness	64
4.	Cohesion	42.5
5.	None	21

Table 3
Overall preferences in paired comparison.

5.2 Corpus study

The above experiment tested readers' preferences among centering constraints. We have also looked for evidence that human authors are sensitive to centering features. Specifically, we address two questions:

Hypothesis	S	N	Probability of chance result
1. V1 preferred to V3	2	28	$p < 0.0005$
2. V3 preferred to V2	9	24	$p > 0.10$
3. V2 preferred to V4	9	25	$p > 0.10$
4. V4 preferred to V5	2	25	$p < 0.0005$
5. V3 preferred to V4	6	26	$p < 0.01$

Table 4
Results of sign test on direct comparisons.

1. Do authors organise and express propositions in a way that promotes fluency, as measured by the four centering features?

2. If so, what relative importance is assigned to the four features? For instance, is salience pursued more strongly than cohesion, or vice-versa?

To investigate these issues we have used a web research tool described by Poesio et al. (2000). The tool uses the GNOME corpus, which comprises two sets of texts, in the museum and pharmacy domains, with centering information fully marked up. Each text is segmented into utterances, and for each utterance the ranked Cfs, Cp, and Cb are identified. In total there are 271 utterances in the museum domain and 397 in the pharmacy domain. The frequencies of the four transitions, as computed by the web research tool with its default parameter settings, are as follows:

	Museum	**Pharmacy**	**Total**
Continuous transitions	187	221	408
Cohesive transitions	84	109	193
Salient transitions	52	95	147
Cheap transitions	78	68	146

At first sight, these figures support the weightings that we have used in trials of the program: continuity seems most important, outranking the other features by a margin of two or three to one. However, this assumes that the actual results achieved by authors are a reliable indicator of what they were trying (unconsciously, of course) to achieve. This might not be true. Perhaps continuous transitions are more frequent because continuity is *easier* to achieve than the other features.

An analogy should make this point clear. In the early stages of a game of darts, players are trying to achieve the highest possible score from each group of three throws. In a match between expert players, most throws are directed at the top segment of the board. Most of this segment scores 20, but a narrow band in the middle of the segment counts as a triple, and scores 60. Three throws into a triple would therefore yield 180 points, the optimum score. However, even among the best players, this score is rarely achieved. On average, players succeed in hitting the triple on about one occasion in three, so that a more typical score for three throws is 100. Now, suppose that we used these statistics as an indication of what the players were trying to achieve. Observing two throws into the 20 region for every one throw into the 60 region, we would infer, wrongly, that players preferred the first outcome to the second.

How can we infer an agent's actual preferences in such a situation? One obvious clue is the likely distribution of outcomes if the agent behaved at random. We might blindfold the player, move the darts board, and then let him throw many darts at random, comparing the number of triple 20s with the number of single 20s. Perhaps in a

million throws he will hit the triple once and the single fifteen times, suggesting that the triple was objectively fifteen times harder. However, complete randomisation is likely to degrade performance so much that we obtain few relevant instances. A better plan would be to randomise more gradually, perturbing the player's performance by small distractions such as a tap on the arm. As the degree of perturbance increases, the ratio of triples to singles will progressively decline, showing that the triple is actually the player's preferred outcome.

Applying this principle to centering, there are two relevant ways in which a text can be perturbed: we can change the order of utterances, or we can change the ranking of Cfs within an utterance. Each perturbance is an operation that shifts some items from their position in the original sequence. To obtain different degrees, we may define a perturbance of order P as follows:

> Given a sequence $X_1..X_N$, take the items in groups of length P, and randomise the order within each group. Thus if $P = 2$, consider first the items $\{X_1, X_2\}$, and order them randomly, so that there is a 50% chance that X_2 will precede X_1. The consider the items $\{X_3, X_4\}$ in the same way, and continue to shuffle locally until reaching the end of the sequence, where the final group might of course contain a single item (or in general, fewer than P items).

Obviously as P increases, it becomes more possible for an item to be moved a long way from its original position, and so the sequence is more thoroughly shuffled.

Our method, in detail, was as follows. A corpus was encoded as a sequence of utterances, each represented schematically by a list of discourse entities ordered according to a salience ranking, so that the first item in the list was the Cp. This encoding ignores some information from the GNOME corpus (specifically, segment boundaries and Cfs carried over from previous utterances), but allows a straightforward computation of perturbed versions. Three *types* of perturbation were considered (utterance order only, Cf ranking only, and both together), and eight *degrees* of perturbation (P varying from 1 to 8). For each perturbed version, the Cp and Cb for each utterance were recomputed, along with the resulting frequencies for the four types of transition (continuous, cohesive, salient, cheap). To smooth out random fluctuations, the computation was performed 10 times for each condition and the results were averaged.

Table 5 shows average percentage frequencies of transitions that respect continuity, cohesion, salience, and cheapness, for both domains (museum and pharmacy), and all types and degrees of perturbation. These frequencies are computed as follows:

Continuity Count the number of transitions (i.e., consecutive pairs) $\{U_i, U_j\}$ in the corpus for which U_j has a Cb. Express this as a percentage of the total number of transitions.

Cohesion Consider *only* the transitions $\{U_i, U_j\}$ for which both U_i and U_j both have Cbs. In other words, ignore transitions for which cohesion cannot be achieved because continuity is not achieved. For the relevant transitions, compute the percentage in which $Cb(U_i) = Cb(U_j)$.

Salience Consider only the transitions $\{U_i, U_j\}$ for which U_j has a Cb. For these transitions, compute the percentage in which $Cb(U_j) = Cp(U_j)$.

Cheapness Consider only the transitions $\{U_i, U_j\}$ for which U_j has a Cb. For these transitions, compute the percentage in which $Cp(U_i) = Cb(U_j)$.

Inspection of table 5 reveals the following:

	Museum domain				Pharmacy domain			
				Perturbing utterance order only				
P	**Con**	**Coh**	**Sal**	**Che**	**Con**	**Coh**	**Sal**	**Che**
1	47	69	44	57	37	79	64	59
2	41	76	47	51	35	76	62	61
3	36	75	50	53	31	80	60	61
4	33	78	47	49	31	78	63	63
5	32	76	47	51	29	77	57	60
6	28	79	47	50	27	80	58	61
7	24	73	48	50	25	79	58	59
8	24	80	49	48	26	78	60	60
r	-0.95*	0.31	0.20	-0.48*	-0.94*	0.04	-0.43*	-0.1
b	-3.2	0.9	0.3	-0.9	-1.7	0.1	-0.7	0.1
				Perturbing Cf ranking only				
P	**Con**	**Coh**	**Sal**	**Che**	**Con**	**Coh**	**Sal**	**Che**
1	47	69	44	57	37	79	64	59
2	47	65	34	44	37	75	47	54
3	47	62	27	34	37	75	44	52
4	47	60	24	32	37	75	42	49
5	47	59	22	32	37	75	40	47
6	47	58	19	29	37	75	41	47
7	47	59	19	27	37	74	39	48
8	47	56	17	28	37	74	40	47
r	0.0	-0.76*	-0.87*	-0.81*	0.0	-0.39*	-0.72*	-0.73*
b	0.0	-1.6	-3.5	-3.6	0.0	-0.5	-2.6	-1.5
				Perturbing both utterance order and Cf ranking				
P	**Con**	**Coh**	**Sal**	**Che**	**Con**	**Coh**	**Sal**	**Che**
1	47	69	44	57	37	79	64	59
2	40	70	34	39	34	75	48	53
3	36	69	30	34	31	79	42	48
4	33	71	27	29	31	82	43	48
5	31	66	25	30	29	76	42	47
6	27	70	22	25	28	78	40	42
7	23	69	23	29	26	78	43	45
8	24	74	25	26	26	80	39	44
r	-0.96*	0.13	-0.7*	-0.75*	-0.9*	0.06	-0.68*	-0.71*
b	-3.3	0.4	-2.5	-3.5	-1.5	0.1	-2.5	-2.0

Con = Continuity, Coh = Cohesion, Sal = Salience, Che = Cheapness
$[r]$ Correlation coefficient
$[b]$ Slope of regression line of frequency F on pertubation P ($F = bP + a$)
$[*]$ Correlations marked with an asterisk are significant at $p < 0.001$ (df=78)
Frequencies for $P = 1$ are those of the original unperturbed texts

Table 5
Average percentage transition frequencies

- Overall, perturbing utterance order and Cf ranking has a clear and progressive effect on all centering features (except cohesion). The frequencies of transitions satisfying continuity, salience and cheapness have a high negative correlation with P, the degree of perturbation, in most conditions. This provides strong

evidence that authors are ordering and expressing utterances in a way that promotes continuity of focus, as measured by these four features.

- The trends within the two corpora (museums and pharmacy) are almost exactly the same. The only evident difference is that continuity was achieved more often in the museum domain than in the pharmacy domain, suggesting that more topic changes occurred in the pharmaceutical documents.

- The effect of perturbing utterance order (but not Cf ranking) is overwhelmingly to disrupt continuity rather than the other features. The correlation between frequency and perturbation for continuity is almost perfect (around -0.95). Cohesion is unaffected, or even improves slightly in the museum domain; salience and cheapness decline little, if at all. When interpreting this data it is useful to consider the regression value b as well as the correlation, since a significant correlation can be achieved even for a slow decline, provided that this decline is regular. For instance, we obtain a correlation of -0.48, significant at the 0.001 level, between perturbation and cheapness, even though the rate of decline is small (only 0.9% for each unit increase in P).

- Perturbing Cf ranking (but not utterance order) cannot affect continuity, by definition. As one might expect, it has a strong negative effect on salience and cheapness; there is also evidence for a small effect on cohesion. Again, note that the b values are much more informative than the significance levels when interpreting the size of the effect — for instance, the highly significant ($p < 0.001$) decline in cohesion for the pharmacy domain is only 0.5% for each unit increase in perturbation.

- Perturbing both utterance order and Cf ranking yields a strong and consistent decline in continuity, salience and cheapness, but has no effect on cohesion. In the museum domain, perturbing utterance order slightly *increases* cohesion, while perturbing Cf ranking slightly *reduces* it; with both perturbations, these possibly artifactual effects, which are far smaller in the pharmacy domain, more or less cancel out.

In summary, we find clear evidence that authors organise texts in a way that promotes continuity, salience and cheapness. For continuity, the correlation with perturbation is over -0.9, and the decline is sizeable as well as steady: 3.3% per unit perturbation for the museum domain, 1.5% for the pharmacy domain. This difference may be due to the higher initial continuity value for the museum domain, in which the texts (so to speak) had more continuity to lose. For salience and cheapness the effect is also strong, and mostly due to perturbation of the Cf ranking. However, we found no clear evidence that authors were pursuing cohesion.

5.3 Discussion and future work

The initial results from the paired comparison study indicate that passages satisfying Cheapness are preferred over other combinations, and that passages satisfying Cohesion on its own are preferred over those where all constraints are violated. These results differ in an interesting way from those of the corpus study. Both studies indicate a preference for Cheap transitions but they do not agree in the relative preferences of Salience and Cohesion. This may indicate a difference between texts that people produce and those they find easier to process; or it may reflect the limitations of this one-off experiment, using fairly simple materials. Further empirical study and more rigorous analysis

will be needed to confirm or refute our current findings. We can conclude however that the results indicate the utility of a system which can produce different output texts by systematically manipulating the parameters of cheapness, salience and cohesion. The next step will be to vary other parameters affecting linear order, such as the putative preferences for right-branching structures and for satellites to precede nuclei, in order to gather evidence on what happens when these constraints conflict with each other or with centering constraints.

Our conclusion from the corpus study can be briefly stated as: utterances are ordered so as to promote continuity, and discourse entities within an utterance are ranked so as to promote salience and cheapness; cohesion is ignored. Of course, by its very nature, such a study tells us only what authors actually do; whether the resulting texts meet the needs of their readers is a separate issue. Cohesion might be a stylistically desirable feature that the run-of-the-mill author is not skilled enough to achieve.

6 Conclusion

We have described a technique for generating texts which will be coherent according to a reasonably faithful interpretation of Centering Theory. Our framework can also incorporate novel features such as Strube and Hahn's "cheapness" principle and a notion of "accessibility" derived from Cristea et al's Veins Theory. As stated in the Introduction, NLG systems need some principled means of deciding on the preferred orderings of clauses and of arguments within clauses, and CT appears a good candidate to provide a basis for these decisions, in tandem with other stylistic considerations. We have reported on a particular implementation in the ICONOCLAST document generation system but the technique can be applied to other NLG systems which perform hierarchical text structuring based on a theory of coherence relations (with additional assumptions as detailed in Section 1):

- for systems which generate a single text plan, CT can determine the most coherent ordering of arguments within clauses;
- for systems which generate multiple text plans, CT can be used to evaluate the different plans as well as determining the optimal realisation of any particular plan.

We have reported empirical studies which provide clear evidence that centering features make a difference to the acceptability of texts and we have demonstrated one way to determine weightings. It may turn out, that different weightings are appropriate for different text genres, or for speech as opposed to "written" text. Our framework will facilitate detailed research into evaluation metrics and will therefore provide a productive research tool in addition to the immediate practical benefit of improving the fluency and readability of generated texts.

Acknowledgements
The essential ideas of this work, namely that centering transitions are epiphenomenal on certain fundamental constraints, that these constraints are suitable for guiding text planning, and that text planning can be formulated as a CSP, were originally presented at the ACL Workshop on Discourse Structure and Reference, 1999, the 12th Amsterdam Colloquium, 1999, and COLING 2000. An earlier version of this paper was presented at INLG 2000, Mitzpe Ramon, Israel. We are grateful to the audiences on those occasions for useful feedback, and also to colleagues on the GNOME project as well as Nikiforos

Karamanis. This work was supported in part by the UK EPSRC under grant references L51126, L77102 (Kibble) and M36960 (Power).

References

Alvarez, Yolanda Vazquez and Mark Huckvale. 2002. The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-Speech Systems. In *Proceedings of International Conference on Speech and Language Processing*, Denver.

Beaver, David. 2003. The optimization of discourse anaphora. *Linguistics and Philosophy*. To appear.

Brennan, Susan, Marilyn Walker Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *Proceedings of 25th ACL*.

Cahill, Lynne. 1999. Lexicalisation in Applied NLG Systems, Technical Report ITRI-99-04. Technical report, Information Technology Research Institute, University of Brighton.

Callaway, Charles B. and James C. Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, Philadelphia.

Cheng, Hua. 2000. Experimenting with the interaction between aggregation and text planning. In *Proceedings of ANLP-NAACL*.

Cristea, Dan, Nancy Ide, and Laurent Romary. 1998. Veins theory: A model of global discourse cohesion and coherence. In *Proceedgins of COLING/ACL'98*, pages 281–285, Montreal.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Henschel, Renate, Hua Cheng, and Massimo Poesio. 2000. Pronominalisation revisited. In *Proceedings of 18th COLING*, Saarbrücken.

Kameyama, Megumi. 1998. Intrasentential centering: A case study. In Marilyn Walker, Aravind Joshi, and Ellen Prince, editors, *Centering Theory in Discourse*. pages 89–112.

Karamanis, Nikiforos. 2001. Exploring entity-based coherence. In *Proceedings of Fourth CLUK*, University of Sheffield.

Karamanis, Nikiforos and Hisar Maruli Manurung. 2002. Stochastic text structuring using the principle of continuity. In *Proceedings of 2nd International Natural Language Generation Conference*, New York.

Kibble, Rodger. 1999. Cb or not Cb? Centering theory applied to NLG. In *Proceedings of ACL workshop on Discourse and Reference Structure*, University of Maryland.

Kibble, Rodger. 2001. A reformulation of rule 2 of Centering Theory. *Computational Linguistics*, 27(4):00–00.

Kibble, Rodger. 2003. Towards the Elimination of Centering Theory. In Ivana Kruijff-Korbayová and Claudia Kosny, editors, *DiaBruck 2003: Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Universität des Saarlandes.

Krahmer, Emiel and Mariet Theune. 2002. Efficient Context-Sensitive Generation of Referring Expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CSLI Publications, pages 223–264.

McCoy, Kathleen and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of ACL workshop on Discourse and Reference Structure*, University of Maryland.

Mittal, Vibhu, Johanna Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467.

Poesio, Massimo, Rosemary Stevenson, Hua Cheng, Barbara di Eugenio, and Janet Hitzeman. 2002. A corpus-based evaluation of centering theory. Technical Report TN-02-01 / CSM-369, Natural Language Engineering Group, University of Essex.

Power, Richard. 2000. Planning texts by constraint satisfaction. In *Proceedings of COLING 2000*.

Power, Richard, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260. Also available as ITRI Technical Report ITRI-03-10.

Prince, Ellen. 1999. How not to mark topics: "topicalization" in English and Yiddish.

Reiter, Ehud. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of 7th International Natural Language Generation Workshop*.

Reiter, Ehud. 2000. Pipelines and size constraints. *Computational Linguistics*, 26(2):251–259.

Reiter, Ehud, Roma Robertson, and Liesl Oman. 2000. Knowledge Acquisition for Natural Language Generation. In *Proceedings of 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Scott, Donia and Clarisse de Souza. 1990. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*.

Strube, Michael and Udo Hahn. 1999. Functional centering – grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Suri, Linda, Kathleen McCoy, and Jonathan DeCristofaro. 1999. A methodology for extending focussing franeworks. *Computational Linguistics*, 25(2):173–194.

van Hentenryck, P. 1989. *Constraint satisfaction in logic programming.* MIT Press, Cambridge, MA.

Walker, Marilyn, Aravind Joshi, and Ellen Prince, editors. 1998. *Centering Theory in Discourse.* Clarendon Press, Oxford.