An Empirical Analysis of Constructing Non-restrictive NP Modifiers to Express Semantic Relations

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

It is not a rare phenomenon for human written text to use non-restrictive NP modifiers to express essential pieces of information or support the situation presented in the main proposition containing the NP, for example, "Private Eye, *which couldn't afford the libel payment*, had been threatened with closure." (from Wall Street Journal) Yet no previous research in NLG investigates this in detail. This paper describes corpus analysis and a psycholinguistic experiment regarding the acceptability of using non-restrictive NP modifiers to express semantic relations that might normally be signalled by *'because'* and *'then'*. The experiment tests several relevant factors and enables us to accept or reject a number of hypotheses. The results are incorporated into an NLG system based on a Genetic Algorithm.

1 Introduction

To produce natural language text, an NLG system must be able to choose among possible paraphrases one that satisfies the highest number of constraints in a certain context. Paraphrases can use various constructions, for example, using nominalisation instead of a clause for event representation. We are particularly interested in the use of non-restrictive (NR) modifiers within a referring expression to express certain semantic relations[1] other than *object-attribute elaboration* (in the sense defined in (Mann and Thompson, 1987)), for instance, causal relations, which are normally expressed by separate clauses connected by cue phrases (Knott, 1996) such as *'because'*.

"A non-restrictive component gives additional information to a head that has already been viewed as unique or as a member of a class that has been independently identified, and therefore is not essential for the identification of the head" (Quirk et al., 1985). This definition can be extended to account for modifiers of not only definite referring expressions, but also definite and indefinite NPs of various types. In this paper, an NR modifier refers to any NP modifying component that is not essential for identifying the object denoted by the head, including all modifiers of an NP that does not intend to identify (e.g. indefinite referring expressions and predicative phrases) (Kronfeld, 1990). Our discussion focuses on definite referring expressions including proper names because of the dominance of such examples in our corpus. However, we would expect no difficulty in applying our observation to other types of NPs.

The semantic roles of NR modifiers, in particular NR clauses, are mentioned in many grammar and linguistics books. Quirk et al. (1985) point out that an NR clause in a referring expression is usually neutral in its semantic role (i.e. it provides descriptive information about its head), but sometimes it can contribute to the semantics of the main clause in a variety of ways. They summarise three types of semantic relations that can be expressed by an NR clause (examples are given in Figure 1):

- *Causal*, where the situation in the main clause is caused by that in the NR clause, e.g. (1a).
- *Temporal*, where the two clauses form a time sequence, e.g. (1b).
- *Circumstantial*, where the NR clause sets a temporal or spatial framework for interpreting the main clause, e.g. (1c).

Halliday (1985) mentions that a subordinate clause can elaborate a part of its primary clause through restating, clarifying, refining or adding a descriptive attribute or comment (see (2) of Figure 1). Halliday's notion of elaboration is much more general than that in other coherence theories like RST (Mann and Thompson, 1987), and the relation expressed in (2) would not be treated as elaboration in most NLG systems.

Similar phenomena were observed from the MUSE corpus[2], a corpus of museum exhibit labels, which

[1] We are concerned with semantic (informational) relations in this paper. Argumentative (intentional) relations are beyond the scope of this paper.

[2] This corpus is collected and annotated for the GNOME project (Poesio, 2000), which aims at developing general algorithms for generating nominal expressions.

(1) a. *He sent ahead the sergeant, **who was the most experienced scout in the company**.*
b. *In 1960 he came to London, **where he has lived ever since**.*
c. *The boy, **who had his satchel trailing behind him**, ran past.*

(2) *Inflation, **which was necessary for the system**, became also lethal.*

(3) *In spite of his **French** name, Martin Carlin was born in Germany and emigrated to Paris to become an ebeniste.*

Figure 1: Examples for NR modifiers contributing to the semantics of the main clauses

describe museum objects on display. For example, in (3) of Figure 1, the modifier *French* is not for identifying the name, but for establishing a concession relation between the main proposition and the subordinate phrase to increase the reader's positive regard for where Martin Carlin was born.

For the convenience of discussion, we define some terminology to be used throughout the paper:

An NR construction/sentence : a sentence that has a main clause and a subordinate NR modifier attached to one of its NPs (e.g. (4b) of Figure 2).

A hypotactic construction/sentence : a sentence that has a main clause and a dependent clause, connected by a cue phrase. This is a common way of expressing semantic relations such as causality (e.g. (4a) of Figure 2). In this syntactic category, we single out a subclass of sentences according to one possible semantic connection between the two clauses. It is defined below.

An elaboration realisation : a type of hypotactic construction where one clause elaborates the semantics of the other. We take cue phrases "as for" or "what is more" to signal elaboration relations[3].

Previous research in NLG mainly focuses on using NR constructions to realise elaboration relations but not other semantic relations (e.g. (Scott and de Souza, 1990) and (Hovy, 1993)). The NR modifier usually adds a descriptive attribute to the object denoted by the head.

The linguistic research suggests for an NLG system the possibility to express certain semantic relations through NR constructions, which is important in two aspects. Firstly, an NR construction gives a more concise alternative realisation for a relation, where the relation is expressed implicitly rather than explicitly and usually more subtly. It does not need cue phrases in most cases, and therefore could avoid using cues too heavily. This could be a better realisation under certain circumstances. Secondly, an NR construction enables a wider range of relations (especially those that are preferred to be expressed implicitly) to be selected for text structuring because the corresponding syntactic option is available.

To understand how to enable an NLG system to generate such modifiers, we are faced with two questions, which are not answered by linguistic research:

1. Can this type of modifier be identified by human subjects, i.e. can humans tell the difference between different NP modifier uses?
2. Under what circumstances can an NR construction be used in substitution of a hypotactic construction without changing the meaning dramatically and how close are the meanings conveyed by the two representations?

An NLG system must come up with some solutions, simple or complex, to these two questions in order to choose among paraphrases. In this paper, we use cue phrases as a signal of semantic relations rather than try to identify the relations directly. We describe systematically controlled experiments aimed at finding out the factors related to the generation of this type of modifier in referring expressions. The result is intended to be reliable enough to be used by NLG systems in generating descriptive text.

2 Corpus annotation

To answer the first question, we annotated the MUSE corpus, from which we have observed three types of modifier uses in an NP:

Firstly, providing properties to uniquely identify the objects or concepts denoted by the NP. Without these modifiers, the NP can denote more than one object/concept or sets of objects/concepts and is ambiguous in its interpretation, e.g. those in (6a). Such modifiers usually appear in phrases headed by the definite article *'the'*, which according to Loebner (1987) has the same meaning in all its uses, including in generic references and predicatives. Modifiers

[3]We acknowledge that these cue phrases are controversial in their semantic interpretations, but not using cue phrases would be even more ambiguous. Besides, our experiment does not heavily depend on these cue phrases.

(4) a. *Private Eye had been threatened with closure because it couldn't afford the libel payment.*

b. *Private Eye, **which couldn't afford the libel payment,** had been threatened with closure.*

(5) a. But P&G contends the new Cheer is a unique formula that also offers an ingredient that prevents colors from fading. *And retailers are expected to embrace the product, because it will take up less shelf space.*

b. *And retailers are expected to embrace the product, **which will take up less shelf space.***

Figure 2: Examples for *inferrability*

in other types of generic references, e.g. indefinites, also belong here.

This type subsumes the modifiers normally considered by the referring expression generation module of an NLG system for uniquely identifying the referents (e.g. (Dale, 1992)).

Secondly, having no effect in constraining a unique or unambiguous concept out of the NP which is either already unique or not required to have a unique interpretation, but being important to the situation presented in the main proposition containing the NP.

This type includes the modifiers described in the previous section and many modifiers in indefinite predicatives, e.g. that in (6b).

Thirdly, providing additional details about the referents of the NP, which functions the same way as the NP without these modifiers, e.g. those in (6c). The effect of such modifiers is usually local to the heads they describe rather than to the main propositions as a whole, which is the main difference between this and the second type of modifier.

This type subsumes the modifiers normally generated by an aggregation module, in particular one using embedding (e.g. (Shaw and McKeown, 1997), (Cheng, 1998)).

(6) a. *the decoration **on this cabinet**; the **best looking** food **I ever saw***

b. *This is a **mighty empty** country.*

c. *the **wide gilt bronze** straps on the coffer fronts and sides; He lived in a **five-room** apartment **in the Faubourg Saint-Antoine**.*

To find out whether the above distinctions make sense to human subjects, we designed an annotation scheme for modifiers in NPs, describing which elements of an NP should be marked as a modifier and how to mark the features for a modifier. Apart from other features, each modifier should be annotated with a pragmatic function feature (PRAGM), which specifies why a modifier is used in an NP. The possible values for this feature are *unique*, *int* and *attr*, corresponding to the three types of modifier uses described above (we will use the value names to refer to the different types of modifier in the rest of this paper). XML was used as the markup language.

We had two trained annotators mark the NP modifiers in the MUSE corpus according to their understanding of the scheme. The agreement between them on the PRAGM feature by means of the Kappa statistic (Carletta, 1996) is .73[4], which means that the distinctions we are trying to make can be identified by human subjects to some extent. The main ambiguity exists between *int* and *attr* modifiers. There seems to be a gradual difference between them and where to draw the line is a bit arbitrary.

In the MUSE corpus annotated so far, 19% of 1078 modifiers in all types of NPs are identified as *int*. So this is not a trivial phenomenon.

3 An experiment

We reduced the size of the problem of when to use an NR construction by focusing on two relations: a causal relation signalled by *'because'* and a temporal relation signalled by *'then'*. The reason for choosing these relations is that the possibilities of expressing them through NR constructions have already been shown by linguists. The two cue phrases are typical for the corresponding relations and can often substitute other cue phrases for the same relations. In the rest of this paper, we will still use the term causal or temporal relation, but what we actually mean is the specific relation signalled by *'because'* or *'then'*.

3.1 Independent variables and hypotheses

From the generation point of view, our question is: given two facts and the semantic relation between them, what extra input do we need for making realisation decisions?

We collected examples of *'because'* sentences from the MUSE corpus and Wall Street Journal source data, and transfered them to NR sentences by hand. Comparing the two constructions, we found some interesting variation. For example, comparing the sentences in Figure 2, we found intuitively that the meanings of (4a) and (4b) are much closer than those of (5a) and (5b). In other words, (4b) can be used in substitution of (4a), whereas (5b) cannot so easily

[4]In (Carletta, 1996), a value of K between .8 and 1 indicates good agreement; a value between .6 and .8 indicates some agreement.

Independent Variables	Levels	
Relation	causal	temporal
Inferrability	strong	weak
Position	initial	final
Order	hypotactic vs. NR	NR vs. hypotactic
Subordination	nuc subordinate	sat subordinate
Cued/NoCue	use cue	not use cue

Table 1: Independent variables and their values

substitute (5a). A similar pattern can be found in a number of other collected sentences.

We claim that it is the degree of inferrability of the relation between the semantics expressed through the two clauses that makes the difference. We define the *inferrability* of a causal/temporal relation as:

> *Given two separate facts, the likelihood of human subjects inferring from their world knowledge that a causal/temporal connection between the facts might plausibly exist.*

In examples (4) and (5), the fact that Private Eye cannot afford the libel payment is very likely to directly cause the closure threaten, whereas a product occupying less space is not usually a cause of it being accepted by retailers according to common sense. Therefore, the two realisations in (4) can be used in substitution of one another whereas those in (5) cannot.

Inferrability is dynamic and user dependent. Given two facts, people with different background knowledge can infer the relation between them with different ease. If a relation is easily recognisable according to general world knowledge, we say that the inferrability of the relation is globally strong, in which case a hypotactic and an NR construction can express the relation almost equally well (if not considering rhetorical effect). Context can also contribute to the inferrability of a relation. A relation not easily recognisable from world knowledge may be identified by a reader with ease as the discourse proceeds. In this case, we say that the inferrability of the relation is locally strong, where the two constructions can express the relation equally well only in a certain context. In this paper, we mainly consider the global aspect of a relation and we will describe how we decided the value of inferrability in the next section.

In Table 1, we summarise the factors (independent variables) that might play a role in the closeness judgement between the semantics of a hypotactic construction and an NR construction. The levels are possible values of these factors. Besides *Relation* and *Inferrability*, *Position* gives the location of the NP that contains the NR modifier. It can be the first (*initial*) or the last (*final*) phrase in a sentence[5]; *Order* gives the order of presentation, a hypotactic sentence to be compared with an NR sentence or vice versa, which is used to balance the influence of cue phrases on human judgement; *Subordination* specifies whether the nucleus or the satellite is realised as an NR clause[6]; and *Cued/NoCue* means using a cue phrase in the NR clause or not, which is only applicable to the temporal relation, for example,

(7) *The health-care services announced the spinoff plan last January,* ***which was <u>then</u> revised in May.***

Based on our observation of human written sentences, we have the following hypotheses:

Hypothesis 1 *For both causal and temporal relations, the inferrability of the relation between the semantics of two facts contributes significantly to the semantic similarities between a hypotactic construction and an NR construction.*

In other words, if the *inferrability* of the relation between the two facts is strong, the semantic relation can be expressed similarly through an NR construction, otherwise, the similarity is significantly reduced.

Hypothesis 2 *For the causal relation, the satellite subordination bears significantly higher similarity in meaning to the hypotactic construction than the nucleus subordination does.*

For example, (4b) would be preferred to "*Private Eye, which had been threatened with closure, couldn't afford the libel payment.*"

Hypothesis 3 *For the temporal relation, both the position of subordination and the use of an appropriate cue phrase in the NR clause make a significant difference to the semantic similarities between a hypotactic and an NR construction.*

This hypothesis prefers Example (7) to the realisation that does not have *'then'*.

[5] In our implementation, we restrict ourselves to sentences with two NPs.

[6] We assume that in the causal relation, the clause bearing *'because'* is always the satellite. Since the temporal relation is a multinuclear relation, this factor does not apply.

Val	Dependent Variables	
	Similarity	Naturalness
6	exactly the same	N/A
5	very similar	natural
4	more similar than different	fairly natural
3	more different than similar	so-so
2	very different	fairly unnatural
1	totally different	unnatural

Table 2: Dependent variables and their values

3.2 The design of the experiment

To assess a semantic similarity, which is thought to be influenced by the independent variables, we use human subjects to judge the following two dependent variables:

Naturalness : how fluent a sentence is on its own.

Similarity : how similar the meanings of two sentences are without considering their naturalness.

The scales of the variables are selected such that all values on the scale have natural verbal descriptions that could be grasped easily by our subjects (see Table 2). Similar rating methods have been described in (Jordan et al., 1993) to compare the output of a machine translation system with that of expert humans.

Since we want to measure different groups of *similarity* judgement based on different *inferrability, order* or *position* levels, a between-groups design (Hatch and Lazaraton, 1991) seems to be most appropriate. The design we used is illustrated in Table 3, where all possible combinations of the independent variables are listed. In the table, *paraphrases* gives the types of alternative sentences each original sentence has. They should be scored by human subjects for their similarities to the original sentences and their naturalness.

We used a method similar to *random selection* to create a stratified random sample. The sample should contain 12 hypotactic sentences and 12 NR sentences, two for each combination of the causal relation and one for each combination of the temporal relation. These numbers were used to obtain as big a sample as possible which could still be judged by human subjects in a relatively short period of time (say less than 30 minutes).

Using cue phrases as the indicators of the semantic relations between clauses, we collected all the sentences containing *'because'* or *'then'* from the Wall Street Journal source data, and went through each of them to pick out those that actually signal the desired relations and can potentially have NR-realisations, i.e. where there is a coreference relation between the two NPs in the two clauses. Sentences containing NR clauses signalled by *', which'* or *', who'* were collected similarly. From these sentences, we randomly selected one by category. If it realised an unused factor combination, it was kept in the sample. This process was repeated until we collected the right number of test items which instantiated all combinations of properties in Table 3.

We asked two subjects to mark the 24 selected items with regard to their inferrability on a five-point scale: 5 for *very likely*, 4 for *quite likely*, 3 for *possibly*, 2 for *even less possibly* and 1 for *unknown*. We took values of 4 and 5 as *strong* and the others as *weak*. The subjects and an author agreed on 19 items, and the author's version was used for the experiment.

For the test items, we manually produced the corresponding paraphrases, which were then put into a questionnaire for human assessment of the two dependent variables for each paraphrase.

3.3 Results

We had ten native English speakers evaluating the similarity and naturalness on the sample.

3.3.1 Similarity

Since the similarity data is ordinal data and departs significantly from a theoretical normal distribution according to One-Sample Komogorov-Smirnov Test, we chose Mann Whitney U, which is a test for comparing two groups on the basis of their ranks above and below the median. The result is summarised in Table 4, with statistically significant items in boldface (taking the conventional .05 p level). The Z scores tell how many standard deviations above or below the mean an observation might be. *Means* gives the means of the similarity scores with respect to the values of the independent variables in Table 1.

For the causal relation, there is a significant difference between the means of similarities of the two groups of different inferrabilities ($P<.0005$). So we have high confidence to accept part of Hypothesis 1. i.e. the strong inferrability of the causal relation between the semantics of two facts makes the semantic similarities between a hypotactic construction and an NR construction significantly higher than the weak case does. In the strong case, the mean of similarity is 4.59, which is close to *very similar*.

We treated *order* as a factor to be balanced and did not expect it to have a significant effect, but it does ($P=.008$). An NR paraphrase shows much higher similarity to its corresponding hypotactic sentence (with a mean of 4.46) than the other way round (with a mean of 3.83), but the difference becomes smaller for the strong inferrability case. This could be because the causal relations expressed in NR sentences generally sound weaker than those in hypotactic sentences and the cue phrase has a big influence on the perceptibility of a relation.

Independent Variables				Paraphrases
Relation	Order	Inferrability	Position	
causal	hypotactic vs. NR sentence	strong	initial	nuc & sat subordination
			final	NR sentence
		weak	initial	nuc & sat subordination
			final	NR sentence
	NR sentence vs. hypotactic	strong	initial	causal & elaboration hypotactic
			final	
		weak	initial	
			final	
temporal	hypotactic vs. NR sentence	strong	initial	cued & not cued NR sentence
			final	
		weak	initial	
			final	
	NR sentence vs. hypotactic	strong	initial	temporal & elaboration hypotactic
			final	
		weak	initial	
			final	

Table 3: A between-groups design

Relation	DependVar	Factors	Means	Z	2-tailed P
causal (160 cases)	Similarity	Inferrability	4.59/3.70	-4.1015	**<.0005**
		Order	4.46/3.83	-2.6400	**.0083**
		Position	4.11/4.18	-.2136	.8308
temporal (80 cases)	Similarity (cued)	Inferrability	4.88/5.00	-.1022	.9086
		Order	5.08/4.80	-1.1756	.2398
		Position	4.80/5.08	-2.0649	**.0389**

Table 4: The output of Mann Whitney U on the similarity data

For the temporal relation, *position* is the only significant factor (P=.0389). So part of Hypothesis 3 is confirmed, that is, the final position subordination makes an NR paraphrase significantly more similar to the corresponding hypotactic construction than the initial position does.

We do not have enough evidence to accept the claim that the inferrability of the temporal relation contributes significantly to the similarity judgement (as in Hypothesis 1). However, when we calculated the similarity mean for the alternative sentences using cue phrases, strong or weak in inferrability, we got 4.94 (*very similar*). Comparing this with that of the strong causal case using the Mann Whitney U test, we get a significance level of 0.0294. This means that we have strong confidence to believe that the similarity mean for the temporal relation if using a cue phrase is significantly higher than that for the strong causal relation. Therefore, the temporal relation can always be realised by an NR construction as long as an appropriate cue phrase is used in the NR clause.

The assumption of normality is also not met by the subset of the data related to Hypothesis 2 and 3 (i.e. the similarity scores for nucleus/satellite subordination paraphrases and cued/nocue paraphrases). We used the Wilcoxon Matched-Pairs Signed-Ranks Test because we were comparing pairs of paraphrases. The result is given in Table 5. We accept the hypothesis that the similarity means of nucleus and satellite subordination are significantly different in the initial position (Hypothesis 2). This confirms the linguistic observation that information of greater importance should be presented in a main position rather than a subordinate position. We can also accept the hypothesis that for the temporal relation, using cue phrases in NR clauses can significantly improve the similarity score of the NR construction (Hypothesis 3).

3.3.2 Naturalness

We used the Mann Whitney U test on *naturalness* with regards to *order*, *inferrability* and *position*, and found no significant connection. Figure 3 shows the distribution of naturalness assessment of the paraphrases for the causal and temporal relation respectively. The majority of the NR constructions are *natural* or *fairly natural*, which suggests that they could be good alternative realisations.

Relation	Paired Variables	Cases	Means	Z value	2-tail Sig
causal	Nuc-similarity/Similarity	40	3.4500/4.4000	-3.4954	.0005
temporal	NoCue/Cued	80	4.3750/4.9375	-3.02	.003

Table 5: The output of the Wilcoxon Matched-Pairs Signed-Ranks Test

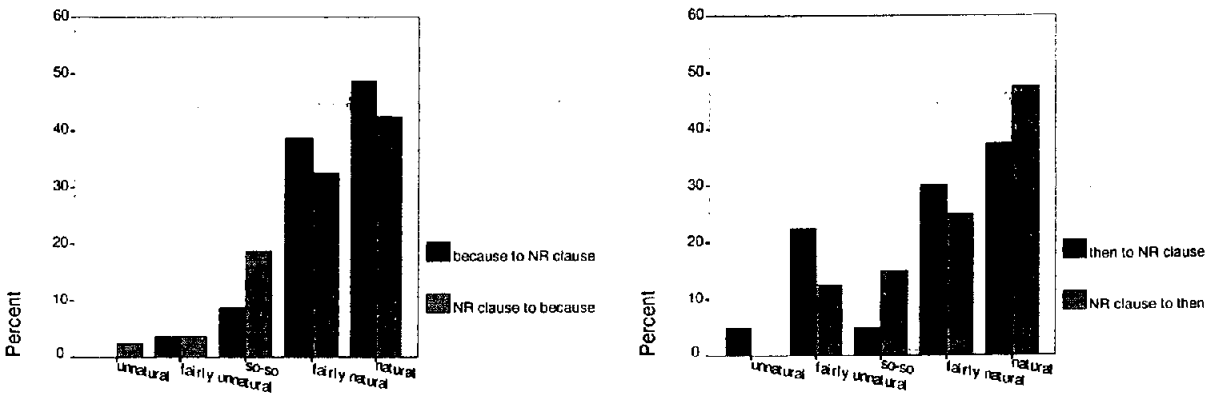

Figure 3: The naturalness of the causal paraphrases (left) and the temporal paraphrases (right)

3.3.3 Summary

We briefly summarise the heuristics drawn from the experiment for expressing the causal and temporal relations with an NR construction. This is an acceptable realisation in the following circumstances:

- the causal relation holds between two facts and the inferrability of the relation is strong, in which case satellite subordination should be used; or
- the temporal relation holds between two facts, in which case a final position subordination and an appropriate cue phrase, like *'then'*, should be used in the NR clause.

We also found that an NR construction can express the causal/temporal relation and the object-attribute elaboration relation at the same time, irrespective of the inferrability of the relation. Generally speaking, a semantic relation expressed by an NR construction sounds weaker than a hypotactic realisation with a cue phrase. Therefore, if a relation is to be emphasised, NR constructions should not be used.

4 Implementing the results in a GA-based text planner

int-modifiers have a mixed character, i.e. like *attr*-modifiers they are not essential for identifying the referents, but like *unique*-modifiers they are not optional. Because of their role in supporting the semantics of the main propositions, the selection of *int*-modifiers should be a part of the text planning process, where a text structure is constructed to fulfill the overall goals for producing the text. However, compared with *unique*-modifiers, *int*-modifiers are less essential for an NP and they can only be added if there are available syntactic slots.

Since embedding deals with *attr*-modifiers at both a content selection and an abstract realisation level, it could coordinate the addition of *int*-modifiers. Therefore, the text planner could consult the embedding module as to whether a property can be realised as an NP modifier, under the constraints from the NP type and the *unique*-modifiers that are already there. In other words, the text planner chooses facts to satisfy certain goals and the embedding process decides if the facts can be realised as NP modifiers in an abstract sense.

We need a generation architecture that allows a certain degree of interaction between text planning, referring expression generation and embedding. So we chose the Genetic Algorithm based text planner described in (Mellish et al., 1998). Their task is, given a set of facts and relations between facts, to produce a legal RST tree using all the facts and some relations. The text planning is basically a two step process. Firstly sequences of facts are generated by applying GA operators, and secondly the rhetorical structure trees built from these sequences are evaluated and the good sequences are kept for producing better offspring.

We extended the text planner by adding a GA operator called *embedding mutation*, which randomly selects two items mentioning a common entity from a sequence and assumes an embedding on them. Embeddings are evaluated together with the other properties an RST tree has. In this way, embedding is performed during text planning. The ultimate score of a tree is the sum of positive and negative scores for all the good and bad properties it bears. Since good embeddings are scored higher, they are kept in the sequences for producing better offspring and are very likely to be included in the final output.

We incorporated the results from the experiment into the GA planner by using them as preferences for evaluating RST trees. We treated *inferrability* as an input to the system. If a good embedding can be formed from two facts connected by an RST relation (i.e. either of the two cases in Section 3.3.3 is satisfied and the required syntactic slot is free), the embedding is scored higher than the hypotactic realisation. However, this emphasis on embedding might not be appropriate. In a real application environment, other communicative intentions should be incorporated to balance the scoring for different realisations. And generally, *inferrability* has to be implemented based on limited domain-dependent knowledge and user configuration.

5 Conclusion and future work

This paper investigates the use of NR modifiers in referring expressions to express certain semantic relations. This is a commonly used strategy by human authors, which has not been explored by an NLG system before. Our experiment shows that when the conditions for *inferrability* etc. are satisfied, certain relations can be expressed through an NR construction as well as a normally used hypotactic construction with little difference in semantics. This facilitates for an NLG system a way of expressing these semantic relations more concisely and subtly which could not be achieved by other means.

Our experiment is restricted in many ways. One possible extension is to use more cue phrases to cover a wider range of cases for each semantic relation. In reality, the application domain should decide which relations need to be tested.

References

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Hua Cheng. 1998. Embedding new information into referring expressions. In *Proceedings of COLING-ACL '98*, pages 1478–1480, Montreal, Canada.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. The MIT Press.

M.A.K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold (Publishers) Ltd., London, UK.

Evelyn Hatch and Anne Lazaraton. 1991. *The Research Manual: Design and Statistics for Applied Linguistics*. Newbury House Publishers.

Eduard Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence 63, Special Issue on Natural Language Processing*, 1.

Pamela Jordan, Bonnie Dorr, and John Benoit. 1993. A first-pass approach for evaluating machine translation systems. *Machine Translation*, 8(1-2):49–58.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh, Edinburgh.

Amichai Kronfeld. 1990. *Reference and Computation*. Studies in Natural Language Processing. Cambridge University Press.

Sebastian Loebner. 1987. Definites. *Journal of Semantics*, 4:279–306.

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.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, Ontario, Canada.

Massimo Poesio. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of LREC*, Athens, May.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Grammar of Contemporary English*. Longman Group Ltd.

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.

James Shaw and Kathleen McKeown. 1997. An architecture for aggregation in text generation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, Poster Session*, Japan.