A corpus analysis of discourse relations for Natural Language Generation

Sandra Williams and Ehud Reiter
Department of Computing Science, University of Aberdeen,
Aberdeen AB24 3UE.
E-mail: swilliam/ereiter@csd.abdn.ac.uk

Abstract

We are developing a Natural Language Generation (NLG) system that generates texts tailored for the reading ability of individual readers. As part of building the system, GIRL (Generator for Individual Reading Levels), we carried out an analysis of the RST Discourse Treebank Corpus to find out how human writers linguistically realise discourse relations. The goal of the analysis was (a) to create a model of the choices that need to be made when realising discourse relations, and (b) to understand how these choices were typically made for "normal" readers, for a variety of discourse relations. We present our results for discourse relations: *concession*, *condition*, *elaboration-additional*, *evaluation*, *example*, *reason* and *restatement*. We discuss the results and how they were used in GIRL.

1. Introduction

The Generator for Individual Reading Levels (GIRL) project is developing a Natural Language Generation (NLG) system that generates feedback reports after a web-based literacy assessment (Williams 2002). The literacy assessment is for adults with poor basic literacy and it was designed by NFER-Nelson for the Target Skills application (Target Skills 2002). It produces a multi-level appraisal of a candidate's literacy skills based on the UK Adult Literacy Core Curriculum (Steeds 2001).

GIRL is being developed with the goal of tailoring output texts to the individual reading skills of users (readers). In working towards this goal, we hope to find out more about generating documents for readers at different reading levels, how to test a system on real users (e.g. both good readers and learner readers), and how to implement reading level decision-making mechanisms as part of the generation process. Also, which decisions produce the most marked impact on the readability of the output texts. We particularly focus on discourse-level choices in GIRL, since earlier projects, e.g. PSET (Devlin et al., 2000), have looked at lexical and syntactic choices for poor readers.

Discourse-level decisions are made in GIRL in a module called the microplanner. The microplanner's input is a tree of discourse relations joining pieces of information. The microplanner plans how the information from the tree will be ordered, whether it will be marked with discourse cue phrases (e.g. 'but', 'if' and 'for example'), and how it will be packed into sentences and punctuated. That is, it makes a number of decisions about how the output text will be realised linguistically. Below are listed a number of different ways in which an *example* discourse relation could be realised.

a) Sometimes you did not pick the right letter. You did not click on the letter 'd'.
b) Sometimes you did not pick the right letter. For example, you did not click on the letter 'd'.
c) Sometimes you did not pick the right letter. You did not, for example, click on the letter 'd'.
d) Sometimes you did not pick the right letter – you did not click on the letter 'd', for example.
e) You did not click on the letter 'd'. Sometimes you did not pick the right letter.
f) Sometimes you did not pick the right letter. For instance, you did not click on the letter 'd'.

These are legal ways of realising an *example* relation that were encountered during this analysis. The discourse relation consists of two text spans, a statement '*sometimes you did not pick the right letter*' and an example '*you did not click on the letter 'd*''. In (a), the ordering of the two text spans is statement-example and each span is a separate sentence (there is a full stop between spans). The version in (b) is the same except that the discourse cue phrase 'for example' has been added before the second text span. In (c) it is positioned mid-second-span. In (d) there is an 'em dash' between spans and the cue phrase is at the end. In (e), the ordering is reversed and there is no cue phrase. (f) is the same as (b) but with 'for instance' rather than 'for example'.

Which version is the most readable? Readability calculators (e.g. Flesch 1949) tell us little about how to produce good readable texts. Flesch's calculator estimates that versions (a) and (e) both score equally as the most 'readable', followed by (b) and (c) equal in second place, then (f) and lastly (d). The PSET project (Devlin et al., 2000) found that lexical selection has an impact on readability, and so cue phrase choice is important. Order of text spans, existence and position of cue phrases and between-span punctuation may also have an impact. Shorter sentences may help poor readers as Flesch's calculator predicts.

It is very important to base the development of GIRL on solid empirical evidence rather than on our own intuitions. There has been very little empirical work to date on what kinds of texts are most appropriate for people with good reading skills and even less on what is appropriate for people with

poor literacy. We were therefore motivated to do our own empirical studies such as the corpus analysis described in this paper.

Aspects of the analysis are unique, since it was customised to the particular requirements of GIRL. Other aspects overlap with the much smaller analysis of Moser and Moore (1996). In this paper, we describe the corpus we used, the method we followed and our results. We go on to discuss how each of the features we analysed has an effect on all the others. We also compare our results with Moser and Moore's. Finally we describe how our corpus analysis results were used to build models for GIRL.

2. Corpus

We first collected a corpus of documents written by literacy tutors for readers at a variety of skill levels, and tried to analyse this to determine our model and the optimal choices for different readers. Unfortunately, this corpus was too small and inconsistent (different tutors wrote in very different ways) to allow us to extract useful information. We hence decided to focus on analysing existing corpora collected by others, even though these corpora do not in general contain texts written for poor readers. Later we adapted the model and choices to poor readers, based of psycholinguistics and experiments, but this is not the subject of this paper.

Initially, we carried out an analysis of discourse cue phrases in the British National Corpus (http://www.hcu.ox.ac.uk/BNC/). That is, we searched the corpus for a number of cue phrases and parts-of-speech. For example, we searched for instances of the cue phrase "but" as a coordinating conjunction. We analysed the first 100 results and on this basis decided that our model would contain the following features: order of discourse text spans; existence and choice of cue phrase(s); position of cue phrase(s); existence and choice of between-text-span punctuation (including sentence-breaking punctuation) and length of first text span. The last two features obviously have an impact on sentence length. Choice of cue phrase has an impact on word length and sentence length. Both sentence length and word length are factors highlighted by readability calculators.

We also attempted to determine choice rules from the BNC analysis, but this was difficult because the BNC only has part-of-speech annotation, not discourse relations annotation. Therefore we had no information about underlying discourse relations. For instance, to find out about the concession relation, a search in the BNC for a cue phrase such as 'but' may give a result containing a few *concession* relations, but it will not identify them nor, of course, give all the *concession* relations. We could not therefore answer some key questions about discourse relation realisation, such as how often authors omitted cue phrases completely, or which cue phrases were preferred for different relations. We obtained the RST Discourse Treebank corpus (RST-DTC) (Carlson, Marcu, and Okurowski, 2002) from the Linguistic Data Consortium (http://www.ldc.upenn.edu/). As far as we are aware, this is the largest publicly available and machine-readable corpus that has been annotated with discourse relations. So it was the best corpus for our purposes. An additional benefit of using this corpus is that the corpus annotators claim not to have been influenced by the presence of discourse cue phrases during their analysis (e-mail from Daniel Marcu).

The RST Discourse Treebank Corpus (RST-DTC) (Carlson, Marcu, and Okurowski, 2002) is based on Rhetorical Structure Theory (RST) (Mann and Thompson 1987). The corpus contains 385 texts from the Wall Street Journal (WSJ) on a number of subjects (e.g. finance, world news and the arts). We analysed part of the TRAINING section. There are 342 files in this section. Each represents a single Wall Street Journal article. The files containing RST annotations are in the format shown in Figure 1.

```
( Root (span 1 156)
  ( Nucleus (span 1 129) (rel2par Topic-Drift)
    ( Nucleus (span 1 40) (rel2par Problem-Solution)
      ( Nucleus (span 1 21) (rel2par span)
        ( Nucleus (leaf 1) (rel2par span) (text _!Kidder, Peabody
                           & Co. is trying to struggle back.<P>_!) )
        ( Satellite (span 2 21) (rel2par circumstance)
          ( Nucleus (span 2 5) (rel2par Contrast)
            ( Nucleus (span 2 4) (rel2par span)
```

Figure 1. A fragment from the RST-DTC annotation file wsj_0604

Segments of text are numbered and nodes are labelled Root, Nucleus or Satellite. The RST tree structure is denoted by embedding text segments and segment spans within parentheses. The figure displays the beginning of one of the WSJ files. It shows the root node spanning over the entire text (segments 1 to 156). Bold font highlights a *circumstance* relation with a single leaf node nucleus

(segment 1), and text: 'Kidder, Peabody...'. The satellite of this relation spans across segments 2 to 21 and contains relations at lower levels in the rhetorical structure tree.

A software tool, RSTtool, is supplied with the corpus for viewing the RST trees. Figure 2 shows a screenshot of the RSTtool that displays part of an RST tree. The tree fragment shows a tri-nuclear *Sequence* relation between three text segments (segment 94: *'He bites it,'*; segment 95: *'scowls'* and segment 96: *'and throws it down'*). The other relation shown is *evaluation-n* with its satellite spanning across segments 94 to 96 and its nucleus is segment 97: *'It's a real dog.'*.

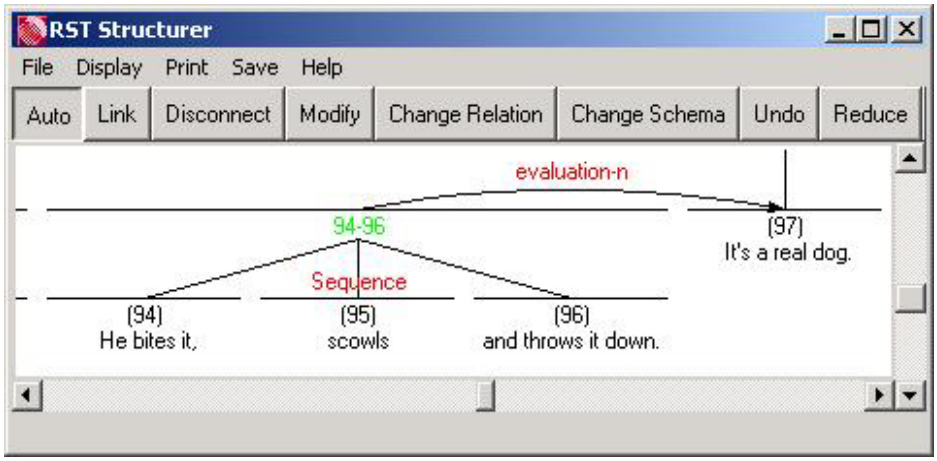

Figure 2. Viewing part of an RST tree using the RSTTool

There are a number of well-documented problems with RST analysis including segmentation of the text, selection of relations, and proliferation of relations (Hovy and Maier 1995). The RST-DTC analysts had to grapple with these and it is not surprising that the corpus suffers from poor inter-annotator agreement (Marcu et al. 1999), as does RST analysis in general. We are not experts in RST analysis, but we did notice some inconsistencies when we were carrying out our own investigations of the corpus, e.g. in segmentation, in choice of relations and in nucleus-satellite identification.

In spite of these problems, we decided to use the RST-DTC in this research project. Investigating the annotated discourse relations within this corpus was the quickest and easiest way to discover how frequently discourse cue phrases are present. Furthermore, it allowed us to investigate how these relations are linguistically realised (text span ordering, cue phrase choice, cue position, etc). The corpus is a useful resource for computational discourse linguistics and it represents a great deal of hard work by its creators.

3. METHOD

The corpus analysis was carried out and checked entirely by the first author. Further checking by independent analysts, would, of course, be desirable, but this was not possible within the scope of the PhD project.

3.1 Choosing discourse relations

Our first task was to choose which of the 60+ discourse relations in the corpus to investigate (the exact number of types of relations present depends on how you count them, e.g. *consequence-n*, *consequence-s* and *Consequence* could be counted as one, two or three). We wanted to analyse the relations most commonly occurring in our small corpus of tutor-written reports, since this is the type of document our NLG system generates. We decided to analyse six relations: *concession*, *condition*, *evaluation* (*evaluation-n* and *evaluation-s*), *example*, *reason* and *restatement*.

The semantics of some relations seem to overlap and we sometimes found it hard to distinguish between them, even though we referred to the definitions in Carlson and Marcu's Discourse Tagging Reference Manual (2002). For instance, in our tutor-written corpus, the tutors gave examples of things the students got right, or wrong. So we chose to investigate the *example* relation. However, another relation, called *evidence*, also seemed close in meaning. Furthermore, tutors often suggested reasons why they thought students had got questions wrong. We decided to investigate the *reason* relation, but we wavered between it and the *cause*, *result* and *evidence* relations. Similarly with the relations *evaluation* and *interpretation*.

We discovered that the relations we chose to investigate were not very common in the TRAINING section of the RST-DTC: *concession*: 1.3%, *condition*: 1.0%, *evaluation*: 0.9%, *example*:

1.6%, *reason* 0.9%, and *restatement*: 0.6%. We later investigated the most common relation of all, *elaboration-additional* (20% of all relations). Other very common relations are *List*: 15.8%, *attribution*: 14.4% and *elaboration-object-attribute*: 12.7%, but these were left for later work.

3.2 Analysing each discourse relation

The corpus analysis proceeded as follows for each discourse relation we investigated. First we searched the annotation files using a simple UNIX grep command for lines containing the relation. From the results of the search, we extracted the RST-DTC file names (e.g. wsj_0603), and the location of the relation (e.g. 'leaf 4' or ' span 102 114'). We produced a table of file names and locations. The table was transferred to an MS Excel worksheet where all analyses for the discourse relation under investigation were recorded. We analysed the first 100 occurrences of each relation.

3.3 Determining the semantic roles of text spans and their ordering

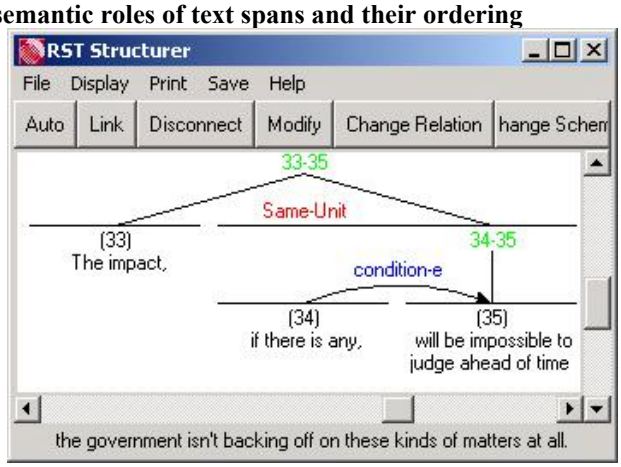

Figure 3. An 'if' text span embedded in a 'then' text span

Next, we opened each file containing the relation under investigation with the RSTtool so that we could view the RST tree and read the text spans of the relation. When viewing the RST tree, we were able to judge (somewhat subjectively) which semantic role each text span of the relation played. For example, in a *condition* relation, one span plays the antecedent role (the one that could potentially have cue phrase 'if') and the other plays the consequent role (the one that could have cue phrase 'then'). To avoid confusion, we simply call these text spans 'if ' and 'then'. We recorded the order of the text spans, e.g. 'if-then' or 'then-if'. Sometimes one span is embedded within another, as in Figure 3 where the 'if' text span, segment 34, (also the satellite) is shown embedded within the 'then' span. Segment 33 and segment 34 are shown linked as the nucleus of this relation by the bi-nuclear 'Same-Unit' keyword. We recorded this embedding as 'th(if)en'.

Note that we did not record nucleus-satellite orderings, but semantic orderings. Some RST relations are in fact defined in terms of semantic roles for nucleus and satellite (Carlson and Marcu 2001), e.g. the condition relation is defined as *"the truth of the proposition associated with the nucleus is a consequence of the fulfilment of the condition in the satellite."* (p.51). However, we found examples in the RST-DTC where these semantic roles were not adhered to and we could not always rely on them.

Nuclearity is also defined in terms of writers' intentions. This is a similar concept to 'intentional subordination' described by Moser and Moore (1996). But RST-DTC annotators were not always able to determine writers' intentions. For instance, in TRAINING there are 47 occurrences of the *cause* relation and 121 of *result*, but there are also 102 occurrences of the multi-nuclear relation *Cause-Result*. This relation is defined in an earlier tagging manual (Marcu 1999) as *"When it is not clear what the intention of the writer is."* (p.24). The proportion of *Cause-Result* relations is relatively high and perhaps this is an indication of a real difficulty the annotators had with this aspect of nuclearity. They could only guess at a writer's intentions, at best. Since we required semantic roles to build our models of discourse realisation for GIRL, we recorded only the semantic role orderings.

3.4 Identifying discourse cue phrases

The next part of our analysis determined whether a discourse cue phrase was present, or not, in the relation being investigated. In most cases, this was fairly clearcut, for instance, we judge that 'if' is a cue phrase in the *condition* relation in Figure 3. But sometimes it was difficult to judge which relation a

cue phrase 'belonged' with. See, for instance, Fig. 4 where we judge that the cue phrase 'and' in segment 35 belongs with the *List* relation, but the cue 'however' in segment 36 belongs with *antithesis* rather than *attribution*. This difficulty often occurred when text spans of the relation spanned across other relations, or vice versa.

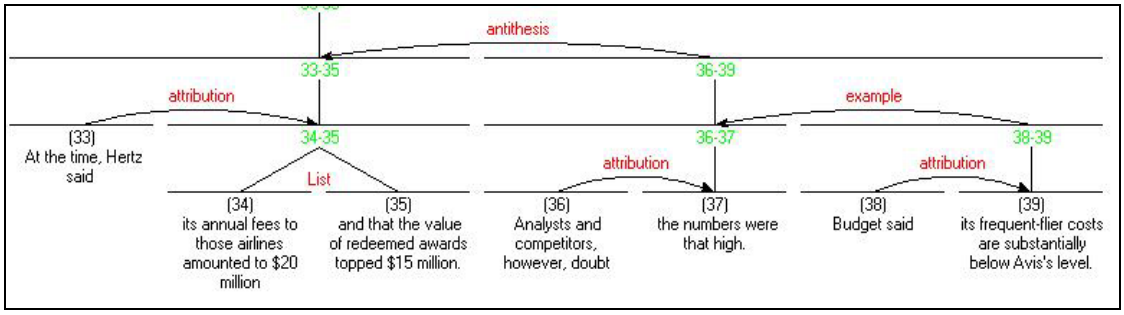

Figure 4. Identifying which relation a discourse cue phrase 'belongs' with.

Another difficulty was judging whether a phrase was, in fact, a discourse cue phrase, or not. We applied the test described by Knott and Dale (1994), i.e. Is the text span containing the potential cue 'incomplete' when elided items have been included and references resolved? We were also helped greatly in identification of cue phrases by using lists that other researchers made available. We believe we have identified some new ones too: for example, 'sure' in an *evaluation* relation with first text span: *'There are, of course, analysts who view the near-panic that briefly swept through investors on Oct. 13 and again on Oct. 24 as momentary lapses of good judgement that have only temporarily undermined a healthy stock market.'*; and second text span: '*Sure, price action is volatile and that's scary, but all-in-all stocks are still a good place to be...'*.

3.5 Recording discourse cue phrase positions and between-text span punctuation

This part of our analysis was not subjective. We recorded the positions of cue phrases, if present, as `before`, `mid-` or `after` a text span. We also recorded between-text-span punctuation, if any. A simple coding scheme was devised to record these details. For instance, `0 . 0` represents two text spans with no cue phrases and a full stop between them. A text span is represented as `0` if there is no cue phrase, `1` if the cue phrase comes before the span, `m` if the cue is mid-span, and `2` if the cue comes after the span. Thus `0 , m` denotes no cue on the first span, a comma between spans and a cue phrase occurring mid-second span. `2 n 1` denotes cue phrases after the first span and before the second span and no between-span punctuation (e.g. '*hurry then if you are late*'). We did not analyse in more detail exactly where mid-span cues occurred (e.g. after the first noun phrase), we leave this for future work. Also, we only recorded the final punctuation mark between spans (e.g. a quote mark followed by a full stop would be recorded simply as a full stop). We also ignored any mid-text-span punctuation.

3.6 Finding the length of the first text span.

We recorded the length of the first text span in words. There are a wide variety of techniques to count words. We tried to be consistent. We counted two hyphenated words (e.g. '*near-panic'*) as two words. We counted people's initials as individual words (e.g. *'A. B. Wilson'* was counted as three words). We counted as one single word: numbers, including monetary amounts (e.g. *'$3,4200'*) and acronyms (e.g. '*USA'*).

4. Results

Here we present the results of our analyses of the RST-DTC.

4.1 Presence of cue phrases

The percentage of instances **without** discourse cue phrases for each relation is shown in Figure 5. Some relations, such as *condition* and *concession*, almost always have cue phrases, so the numbers of examples with no cues are very small. Other relations, such as *restatement*, almost never have cue phrases (99% of cases found). Other relations fall between these extremes.

Our results for presence of cue phrases agree fairly well with Moser and Moore's (1996). They found 133 discourse relations with cue phrases out of a total of 286 relations in their corpus (47%). We found 310 relations with cue phrases out of a total of 700 relations (44%). Our result is not entirely representative of the WSJ texts since we isolated seven discourse relations for our study. However, we chose to include elaboration-additional because it is the most common relation and only

24% of instances of this relation had cues. So 44% is an overestimate. The slight differences might be due to differences in genre of the two corpora. Moser and Moore's corpus consists of on-line tutor feedback and this is actually much closer to the kind of texts that GIRL generates, although, like RST-DTC it does not contain text written for poor readers.

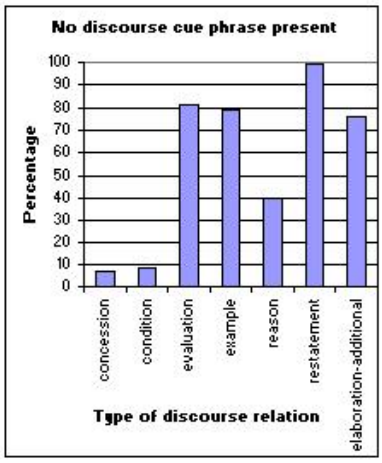

Figure 5. No cue phrases present

4.2 Writers' cue phrase preferences

We compared the actual cue phrases chosen by authors to mark each relation. Table 1 shows all the cue phrases we found and in which of the seven relations investigated they occurred.

N	cue phrase	r
1	accordingly	cc
1	after all	re
1	also	el
11	although	cc
1	and	cc
4	and	el
8	and	ev
4	and	re
1	and for that reason	re
1	and therefore	re
1	as a result	re
2	as long as	cd
2	as soon as	cd
40	because	re
35	but	cc
2	but	el
2	but	ev
1	but	re
8	despite	cc

n	cue phrase	r
1	even as	cc
1	even before	ex
4	even if	cd
3	even though	cc
2	except	el
14	for example	ex
3	for instance	ex
2	given	cd
3	however	cc
1	however	el
2	however	ev
62	If	cd
3	if ... then	cd
1	In fact	el
1	indeed	re
1	instead	re
1	just as soon as	cd
1	mainly because	re
1	now	el

n	cue phrase	r
1	once	cd
1	only if	cd
1	or	rs
2	particularly because	re
1	particularly since	re
1	partly because	re
1	rather	cc
2	regardless	cc
3	should	cd
1	so	ev
1	so	re
1	so far	ev
5	still	cc
2	such as	ex
1	sure	ev
3	that	el
7	though	cc
5	unless	cd
5	until	cd

n	cue phrase	r
1	whatever	el
1	when	cc
2	when	cd
1	when	ex
3	when	re
1	whereas	cc
1	whether	cc
4	which	el
2	which	ev
8	while	cc
4	who	el
1	with	ev
2	without	cd
1	yet	cc
1	yet	ev
390	no cue phrase	
700	**TOTAL**	

Table 1. Cue phrases used in 700 relations, where n=number, r=relation, cc=concession, cd=condition, el=elaboration-additional, ev=evaluation, ex=example, re=reason and rs=restatement.

Table 1 shows marked differences in the range of cue phrases used. For instance, *concession* had 17, although many of these occurred only once. *Restatement* had only one occurrence of a cue phrase ('or') in the 100 cases investigated. The cue phrase 'if' has a high frequency (62) in Table 1 because writers choose to mark over 90% of condition relations with cues, and the cue most commonly chosen is 'if'.

Of the relations with many different cue phrases, usually one or two cue phrases occur with much greater frequency than others. For instance 'but' most often marks a *concession* relation, 'if' a *condition* relation, and 'because' a *reason* relation.

4.3 Order of text spans

Figure 6 shows the semantic ordering of text spans. We found *restatement* **always** occurs in an order where a statement is followed by a restatement (statement-restatement), whilst *example*, and *elaboration-additional* almost always occur in orderings statement-example and statement-elaboration. *Concession*, *evaluation* and *reason* all have one preferred ordering, whereas *condition* occurs in both orders with almost equal frequencies. Embedding is relatively rare. In these seven relations, it is most

common in *condition*. In TRAINING as a whole, the frequency of embedded relations is 13%, which is much higher. Embedding is likely to have an impact on readability and it should be investigated further. Note that we call the spans in *concession* 'but' and 'though' where 'though' denotes the span that could be marked with cue phrases 'though', 'although', etc. and 'but' denotes the other (see section 3.3).

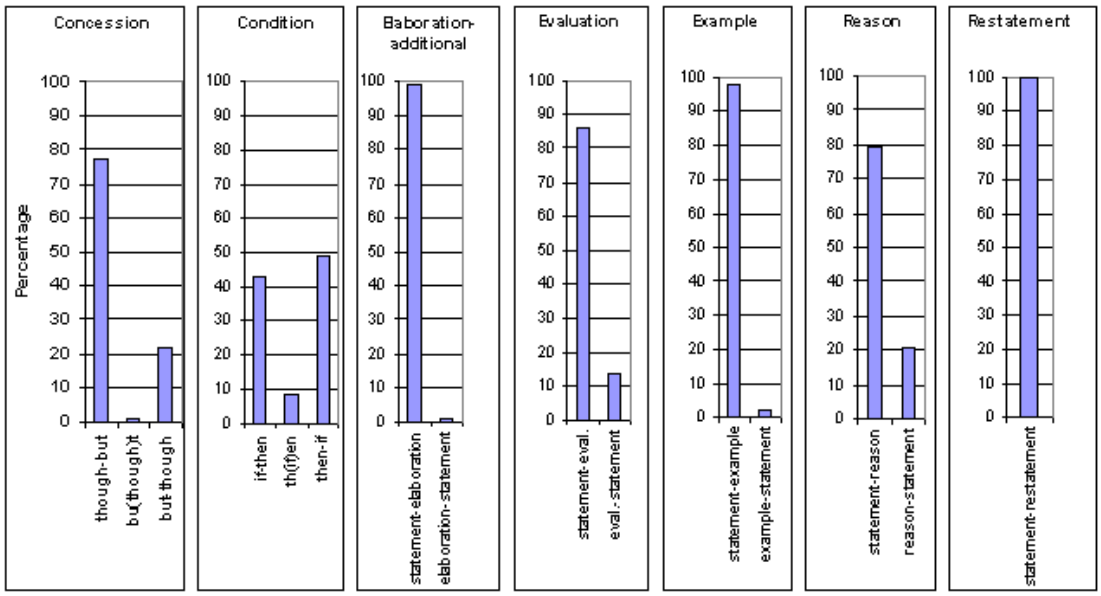

Figure 6. Results for text span ordering for seven discourse relations

4.4 Positions of cue phrases and between-span punctuation

Figure 7 shows the positions of cue phrases (before, mid-, or after the first or the second text span) for each relation and between-span punctuation. The coding scheme is described in section 3.5. We found one relation, *restatement*, was actually more commonly marked by an open parenthesis, than by a cue phrase.

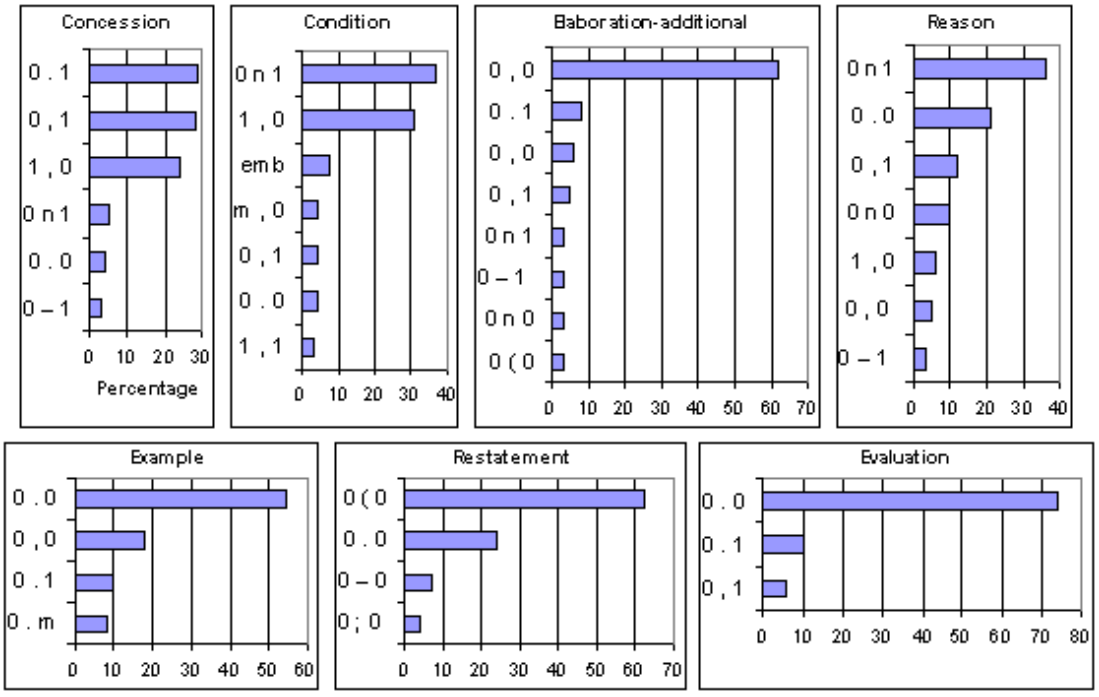

Figure 7. Positions of discourse cue phrase(s) and between-text-span punctuation

C*ondition*, has a distinct correlation between order of text spans, cue phrase position and punctuation. The code `0 n 1` (e.g. 'hurry if you are late') accounts for 37% of all condition relations and `1 , 0` (e.g. 'if you are late, hurry') accounts for a further 31%. So, if a cue phrase is present before the second

span, a comma is not required and vice versa. This correlation appears to happen to a slightly lesser extent in *reason* where `0 n 1` occurs with frequency 36% and it occurs too in *concession*.

4.5 Length of first text span

Figure 8 shows the results for the length of first text spans for all seven relations. The curved lines drawn over the histograms indicate different trends. The relations fall into three categories showing similar trends. *Concession* and *example* have fairly flat profiles, indicating that these relations can have first text spans of any length. *Condition*, *reason* and *restatement* have quite short first text spans with the curves reaching a maximum at 6-10 words. *Elaboration-additional* and *evaluation* reach their maxima above 30 words. The reason why their first text spans are so long is that these two relations often have left-branching RST tree structures. That is, their first text spans tend to stretch across many relations at lower levels in the RST tree.

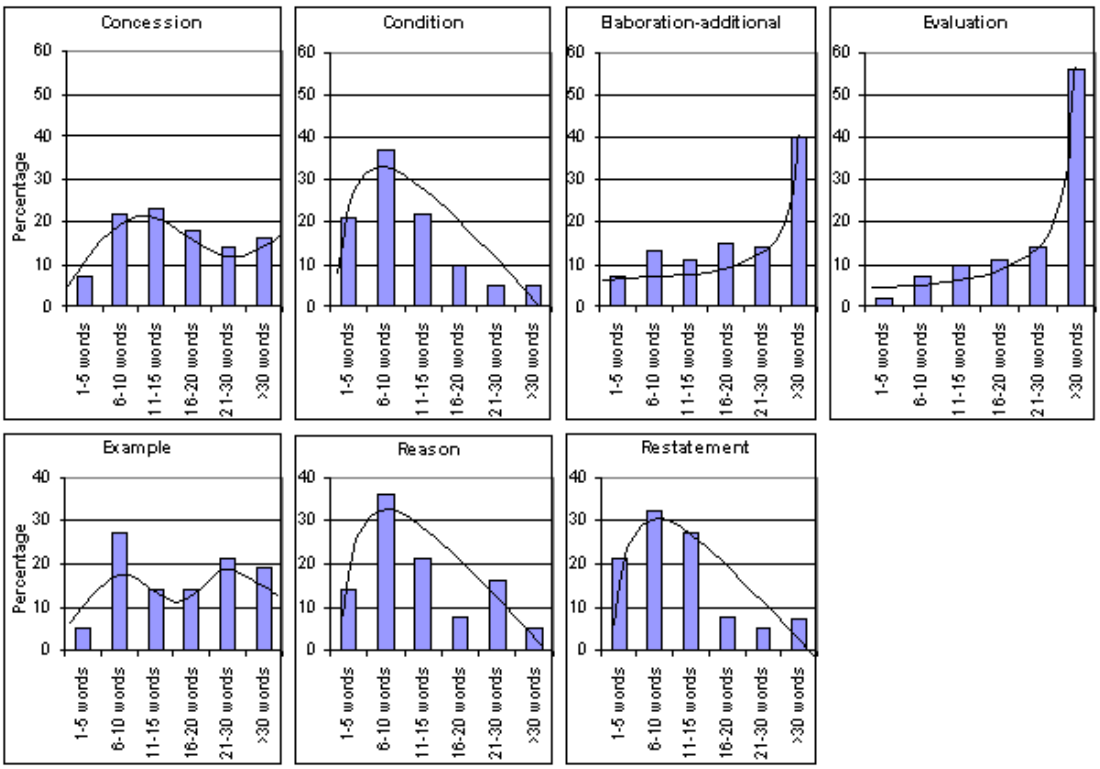

Figure 8. Length in words of first text spans for seven discourse relations

5. Cross-correlations

We have presented our results for some individual features involved in the linguistic realisation of discourse relations, i.e. first text span length, span order, existence and choice of between-span punctuation, existence and choice of cue phrases and cue phrase position. But what we were really interested in is the impact of each feature on all the others, so that we could build models for GIRL. It is easiest to illustrate what we mean by Table 2. This table is a matrix of rules derived from results for just one of the relations: *concession*. The effect of one feature on any other can be seen by finding one feature on the left-hand side of one of the rows, then locating the other feature in one of the column headings (or vice versa) and reading the table cell located where the row and column intersect. The rules shown in this matrix are simplified with the less frequent results omitted.

Some correlations are as we would intuitively expect. For instance, the impact of first span length on punctuation is that with a shorter span a comma is preferred (1-10words->comma in the table), whereas with a longer span a full stop is preferred. Other effects are less obvious, e.g. the cue choice rules.

Moser and Moore (1996) use a theory of discourse relations with 'core' and 'contributor'. These are somewhat similar to RST's nucleus and satellite (where core is roughly equivalent to nucleus and contributor to satellite). They discovered that a cue phrase **never** occurs on the core if it is ordered first. In our cross-correlation for *concession* in Table 2, the summary for order vs. cue position says that in 'but-though' order (see section 4.3 for an explanation), the cue cannot go on the first text span, only on the second. The *concession* relation is defined semantically in RST and what we call the 'but'

text span is the nucleus. We did not find any counter examples in our semantic analysis and this result confirms that the RST-DTC annotators correctly assigned the nucleus and satellite for *concession*. Also our result agrees with Moser and Moore's result, since our but-though ordering is equivalent to nucleus-satellite ordering and hence in this order, a cue phrase can only go on the satellite.

	LENGTH	ORDER	PUNCTUATION	CUE CHOICE	CUE POSITION
LENGTH		No correlation	**1-10words** -> comma **11-20words**->comma, full stop **>20words** -> full stop	**1—10**->but, despite, while **11-20**->but, though, although, while **>20**->but, still, no cue	**1-20words**-> before either **>20words**->before 2nd
	ORDER		**though-but**->comma, full stop **but-though**->comma, none	**though-but**->no cue, but, although, despite, still, while, though **but-though**->although, even though, despite, though	**though-but**->before either **but-though**->before 2nd
		PUNCTUATION		**comma**->although, but, despite, though, while **full stop**->but, still, no cue	**comma**->before either, **full stop**->before 1st, no cue **none**->before 2nd
			CUE CHOICE		**although**->before either **but**->before 2nd **despite**->before either **still**->before 2nd **though**->before either **while**->before either

Table 2. A simplified cross-correlation matrix showing the most common results for *concession.*

6. Models for NLG

Instead of a matrix of rules for each relation, such as the one for *concession* in Table 2, we constructed a decision tree and sets of rules using machine learning techniques similar to Di Eugenio, Moore and Paolucci (1997). We fed in a set of feature-values for each instance of a relation (e.g. for a condition relation this might be: first span length = 8, order = if-then, cue = if, position = before-first, punctuation = comma). Then the machine learning algorithm gave us a decision tree from which we derived a set of rules for the microplanner. It can then work out from the lengths of input text spans: their ordering, the existence, selection and position of cue phrase(s) and existence and selection of between-span punctuation. We built a model in this manner for each discourse relation. Our current microplanner uses these models, as described in the introduction.

Although it is not the topic of this paper, we have subsequently adapted the choices specified by the models derived from this corpus analysis to be more appropriate for learner readers. We based these adaptions on psycholinguistic findings and on our own experiments (Williams and Reiter 2003). We are currently testing readability (reading speed and comprehension) for various readers on texts generated using different choice profiles.

7. Conclusions

This paper has described an analysis of the RST Discourse Treebank Corpus with the goal of acquiring rules for realising discourse relations. Our results show that discourse relations are commonly used in *concession*, *condition* and *reason* whilst they are rarely used in *restatement*, *evaluation*, *example*, and *elaboration-additional*. Human writers most commonly use the discourse cue phrases 'but' in *concession*, 'if' in *condition*, 'because' in *reason* and 'for example' in *example*. They tend to prefer one text span ordering over the other for all relations except *condition*. *Evaluation* and *example* are most commonly realised with no cue phrase and text spans in separate sentences. *Elaboration-additional* most commonly has no cue phrase and a comma between segments. *Restatement* is most commonly marked by punctuation,'('. We found that correlations between order of spans, position of cue phrase and punctuation were indicated in *condition*, *reason* and *concession*. We discussed a correlation matrix for concession and showed examples of derived rules.

This corpus analysis allowed us to achieve our goal of creating models in GIRL for the choices that need to be made when realising discourse relations and how these choices are typically made for "normal" readers.

Acknowledgements
An EPSRC studentship award supported this work. We would like to thank Alistair Knott, Daniel Marcu, Chrysanne Di Marco and Robert Mercer for lists of discourse cue phrases; everyone involved in the creation of the RST-DTC; Somayajulu Sripada and Judy Delin for helpful comments.

References
Carlson L, Marcu D, Okurowski M, 2002 Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Kuppevelt and Smith (eds) *Current Directions in Discourse and Dialogue,* Kluwer.

Carlson L and Marcu D 2001 Discourse Tagging Manual. Information Sciences Institute, University of Southern California. ISI Tech Report ISI-TR-545.

Devlin S, Canning Y, Tait J, Carroll J, Minnen G and Pearce D 2000 An AAC aid for aphasic people with reading difficulties. *Proceeding of the 9th Biennial Conference of the International Society for Augmentative and Alternative Communication*. Washington D.C.

Di Eugenio B, Moore J D and Paolucci M 1997 Learning Features that Predict Cue Usage , in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

Flesch R [1949] *The Art of Readable Writing.* Harper. USA.

Hovy E H and Maier E 1995 *Parsimonious or Profligate: How Many and Which Discourse Structure Relations?* Information Sciences Institute, University of Southern California. Unpublished manuscript.

Knott A and Dale R 1994 Using Linguistic Phenomena to Motivate a Set of Coherence Relations. *Discourse Processes*, Volume 18, Number 1, pp. 35-62.

Mann W C and Thompson S A 1987 Rhetorical Structure Theory: A Theory of Text Organization. Information Sciences Institute (ISI) Reprint no. ISI/RS-87-190, University of Southern California.

Marcu D 1999 Instructions for Manually Annotating the Discourse Structure of Texts. Information Sciences Institute, University of Southern California.

Marcu D, Romera M, and Amorrortu E 1999 Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. *Workshop on Levels of Representation in Discourse.*

Moser M and Moore J 1996 On the correlation of cues with discourse structure: results from a corpus study. Unpublished manuscript.

Steeds A (ed) 2001 Adult Literacy core curriculum including Spoken Communication. Produced by Cambridge Training and Development Ltd. for The Basic Skills Agency. ISBN 1-85990-127-1

Target Skills Literacy Assessment. 2002. Published by the Basic Skills Agency, in association with Cambridge Training and Development Ltd. and NFER-Nelson Ltd.

Williams S H and Reiter E 2003 Experiments with discourse-level choices and readability. To appear in *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest.

Williams S H 2002 Natural language generation of discourse connectives for different reading levels. In *Proceedings of the 5th Annual CLUK Research Colloquium*, University of Leeds.