Generating Embedded Discourse Markers from Rhetorical Structure

Richard Power, Christine Doran and Donia Scott*
Information Technology Research Institute
University of Brighton
Lewes Road, Brighton BN1 6LR England
firstname.lastname@itri.brighton.ac.uk

Abstract

To understand a discourse, the reader needs to recover the relations between the discourse elements as intended by the writer. Writers can, and very often do, help the reader along by providing explicit lexical signals of the intended discourse relations through the use of lexicalised discourse markers. In this paper we present a novel approach for generating texts containing multiple, embedded rhetorical relations, each of which is lexically marked. Our analysis integrates syntactic factors, ideas from rhetorical structure theory, notions of text-grammar, and findings from psycholinguistic research into a unified, feature-based account. The current implementation allows us to generate several good variants of text structures containing multiple discourse relations, while blocking dispreferred variants.

1 Introduction

To understand a discourse, the reader needs to recover the relations between the discourse elements as intended by the writer. Writers can, and very often do, help the reader along by providing explicit lexical signals of the intended discourse relations through the use of *discourse markers*[1] such as 'although', 'nonetheless', 'in order to' and 'on the other hand'. Di Eugenio et al. (1997) and Grote and Stede (1998) emphasise the three types of decisions which need to be made in generating appropriate discourse markers: *occurrence*, whether to generate a marker; *placement*, where to place the marker; and *selection*, which marker to use. This work concentrates on *placement* and the interaction between *placement* and *selection*.

The approach here differs from previous work in presenting a more formal analysis of the generation of discourse markers which combines empirical, linguistic and psycholinguistic factors. It takes into consideration the constraints imposed by syntax, semantics and text structure to generate texts like these:

(1) Elixir has no significant side-effects. **However**, the medicine is for you, **so** never give it to other patients.

(2) Elixir has no significant side-effects, **but since** the medicine is for you, never give it to other patients.

while avoiding texts like these:

*This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant L77102.

[1] Also referred to variously as *clue words* (Reichman, 1981), *cue phrases* (Grosz and Sidner, 1986), *clue phrases* (Cohen, 1987), *rhetorical markers* (Scott and de Souza, 1990), *sentence and clausal connectives* (Knott and Mellish, 1996), *discourse cues* (Di Eugenio et al., 1997), and *discourse connectives* (Webber et al., 1999).

(3) #Elixir has no significant side-effects. **But** the medicine is for you, **consequently**, never give it to other patients.

(4) #Elixir has no significant side-effects. **Since** the medicine is for you, never give it to other patients, **however**.

What is particularly interesting about these texts, which are representative of our target data, is that they contain multiple, embedded rhetorical relations. There is little previous work on multiple discourse markers, in particular in marking embedded discourse relations. Webber et al. (1999) allow for multiple discourse markers in their LTAG discourse representations, but these are between single pairs of propositions. In contrast, the cases which we are considering have only a single discourse relation between a pair of propositions, but either of the pair may contain a second complex discourse within it. To our knowledge, there is no previous work on generating multiple discourse markers governing spans larger than the sentence.[2]

We know that there are usually several ways to realise a given rhetorical relation, so it follows that if we have embedded relations and want to mark them all explicitly, the range of options will be even larger. Our goal in the work described here is to efficiently generate only the licensed combinations of discourse markers for communicating the rhetorical relations. In Section 2, we present our approach to generating discourse markers, and in particular to generating appropriate discourse relations for EMBEDDED discourse relations. While our machinery may seem excessive for generating single discourse markers, the strengths of our approach become evident in the ease with which we are able to handle multiple discourse markers. Section 3 describes the implementation of this approach in the domain of Patient Information Leaflets (the inserts patients receive with medicines).

2 A feature-based treatment of discourse markers

Our analysis integrates syntactic factors, ideas from rhetorical structure theory, notions of text-grammar, and findings from psycholinguistic research.

We utilise the insights from Scott and Souza (1990) on generating texts which efficiently and accurately convey the intended rhetorical relations, and which follow established psycholinguistic findings on ease of processing. We represent text structure as a fairly standard rhetorical structure tree in which the non-terminal nodes are RST relations (Mann and Thompson, 1988), the terminal nodes are *propositions*[3], and the arcs are labelled *nucleus* or *satellite*. Various feature annotations on the nodes of the tree control the ordering of clauses, the choice and placement of discourse markers, and the placement of sentence boundaries. Certain constraints are essential for generating comprehensible and coherent texts, while others are simply stylistic preferences. This section will focus on those which are essential, i.e. those which form the heart of our analysis, while section 3 will describe some of the optional constraints, which may be considered implementation choices.

A somewhat incidental classification in Knott's thesis (Knott, 1996) has proven extremely useful in our analysis. In his Appendix A, he lists all of the discourse markers considered in his corpus analysis, along with their "syntactic category": conj-adverb, subordinator, coordinator, prepositional phrase or phrase with sentential complement (we are calling the conj-adverb class *parenthetical*). The first three of these categories appear frequently in our domain, and are what we will concentrate on here.

What is useful about this classification is that it gives us the positions in which each discourse marker may appear. A discourse marker with the category *parenthetical* (e.g., 'however') can occur in several positions, but only in the linearly second (simple or complex) sub-unit of the

[2]See (Rösner and Stede, 1992) for a schema-based approach which allows multiple discourse markers in a single sentence.

[3]By *proposition* here, we mean the meaning of a rhetorically simple text unit (i.e. one that is not analysed further at a rhetorical level).

relation; a *coordinator* (e.g., 'but') must be expressed at the beginning of the second sub-unit; and a *subordinator* (e.g., 'although') may be expressed at the beginning of either the first sub-unit or the second. This means that with *subordinator* discourse markers, the linked spans can occur in either order. For *parenthetical* or *coordinator* discourse markers, the order of the spans is determined by whether the expression is realised on the nucleus or the satellite of the relation.

Based on our analysis of the possible combinations of discourse marker classes, we find that discourse markers have a "scope" ordering that falls out from their syntactic properties. Because *parenthetical* markers have the freest syntax, they can often be used higher up in the rhetorical tree, outscoping *subordinator* and *coordinator* markers. For instance, given the rhetorical structure in Figure 1, where `concession` dominates `justify`, the text in (5) is preferable to that in (6).

(5) Elixir has no significant side-effects. **However, since** the medicine is for you, never give Elixir to other patients.

(6) #**Although** Elixir has no significant side-effects, the medicine is for you, **consequently**, never give Elixir to other patients.

In the first passage, the selected *parenthetical* marker for `concession` ('however') outscopes the *subordinator* marker chosen for `justify` ('since'), so the rhetorical structure is expressed clearly. In the second passage, a *subordinator* marker for `concession` ('although') is outscoped by a *parenthetical* marker for `justify` ('consequently'), and the result sounds awkward. It would sound better with the last proposition in a separate sentence:

(7) **Although** Elixir has no significant side-effects, the medicine is for you. **Consequently**, never give Elixir to other patients.

but this passage expresses a different rhetorical structure, with `justify` dominating `concession`.

There are also text-grammar (Nunberg, 1990) constraints imposed by the discourse markers. We have four levels in the text grammar: `phrase` (less than a full orthographic sentence), `sentence` (orthographic sentence), `paragraph` and `section`. For *subordinator* discourse markers, the two linked spans must both have LEVEL = `phrase`. For *coordinator* discourse markers it is preferable (in formal writing styles) for both linked spans to have LEVEL = `phrase`. For *parenthetical* discourse markers it is preferable for the linked spans to have LEVEL > `phrase`. To illustrate these constraints, here are three passages that violate them:

(8) #Never give Elixir to other patients. **Although** it has no significant side-effects.

(9) #Elixir has no significant side-effects. **But** never give it to other patients.

(10) #Elixir has no significant side-effects, **however**, never give it to other patients.

Figures 2 and 3 show how we integrate these various constraints to produce two possible specifications of the simple rhetorical structure shown in Figure 1, corresponding to the texts in examples (11) and (12) respectively.

(11) Elixir has no significant side-effects. **However**, the medicine is for you, **so** never give Elixir to other patients.

(12) Elixir has no significant side-effects, **but since** the medicine is for you, never give Elixir to other patients.

The initial rhetorical structure has only relations annotated, along with the nucleus and satellite roles of the participation propositions. It is not until the specific discourse markers are selected that the order of clauses (POSITION) and text-grammar category (LEVEL) are fixed.

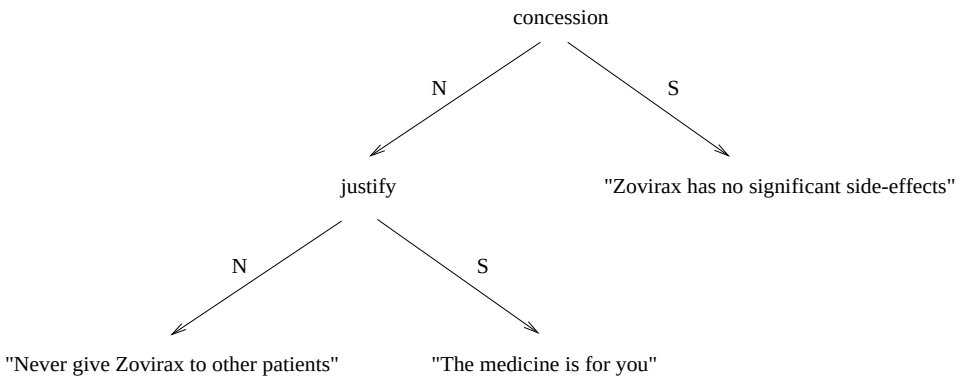

Figure 1: Example of rhetorical structure

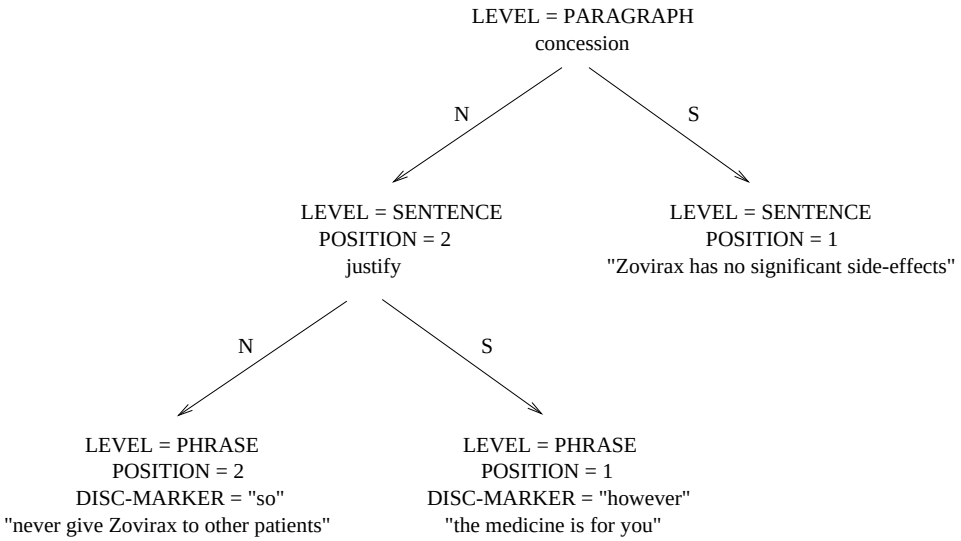

Figure 2: Informal text structure

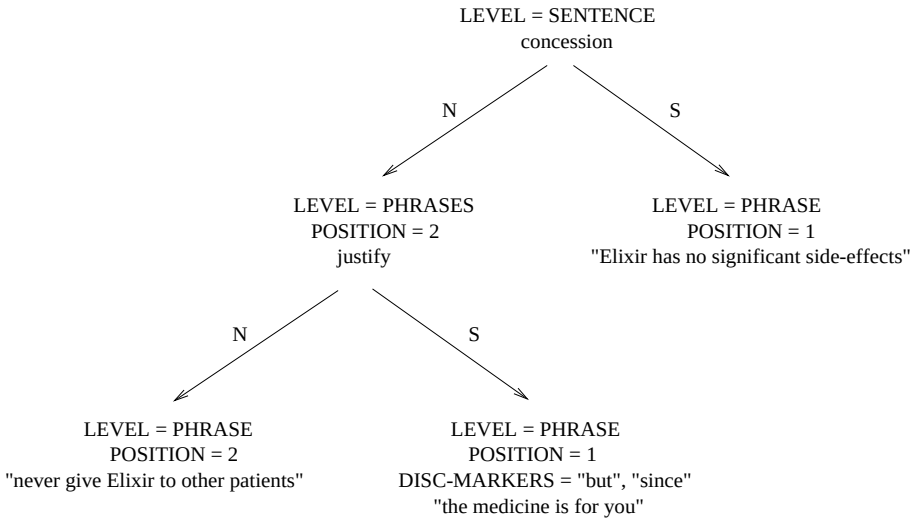

Figure 3: Another informal text structure

3 Implementation

This analysis has been implemented using constraint logic programming, in a generation system for Patient Information Leaflets (PILs). We are currently able to generate all licensed variants of text structures containing multiple discourse relations, while blocking dispreferred variants. The texts generated will vary along a number of dimensions, based on the interaction of the constraints already noted: order of propositions, discourse markers and placement of sentence boundaries.

We distinguish two kinds of text-structure units, corresponding to the terminal and non-terminal nodes in the figures: *simple units* and *complex units*. These share most textual features, but differ in their semantic features. The meaning of a complex unit is represented by the features RELATION, NUCLEUS and SATELLITE; in the figures, the relation is the rhetorical label on the node, while the nucleus and satellite are the units pointed to by the outgoing arcs. The meaning of a simple unit is represented by the feature PROPOSITION, whose value (for present purposes) will be regarded merely as a string of words.

The full list of features and values used in simple and complex units is shown in Table 1.

simple or complex unit	**complex unit only**
NUMBER: single, multiple	DISC-MARKER: *complex feature*
POSITION: 1, 2	RELATION: concession, justify, etc.
CUE-STORE: *passes values of higher nodes*	
LEVEL: phrase, sentence, paragraph, section	

Table 1: Features available on nodes of text structure tree, and their possible values

In the basic rhetorical structure (Figure 1), nucleus and satellite are not assigned a linear order. The POSITION feature specifies the order of the unit in relation to its sister, once that has been fixed. The CUE-STORE feature serves to pass information about discourse markers down the tree from the (non-terminal) node where they are introduced to the (terminal) node where they are expressed. Where they are realized and the POSITION options available both depend upon the

syntactic categories of the specific discourse markers selected for each relation. NUMBER simply indicates whether the text unit is made up of one or several sub-units at the same LEVEL. For instance, a `paragraph` might have as one daughter a relation composed of two complete `sentences`; thus, that daughter would have LEVEL=`sentence` and NUMBER=`multiple`. The DISCOURSE MARKER feature has four sub-features defining its meaning, wording, syntax, and placement. These features and values are: RELATION (the rhetorical relation), PHRASE (the lexical realisation), SYNTAX (subordinator, coordinator or parenthetical)[4] and LOCUS (nucleus or satellite).

As examples, here are three discourse marker definitions for the `concession` relation.

- RELATION `concession`
  LOCUS `satellite`
  SYNTAX `subordinator`
  PHRASE 'although'
- RELATION `concession`
  LOCUS `nucleus`
  SYNTAX `coordinator`
  PHRASE 'but'
- RELATION `concession`
  LOCUS `nucleus`
  SYNTAX `parenthetical`
  PHRASE 'however'

Figure 4 illustrates these features in a more complete version of the text structure shown informally in Figure 3, with discourse markers selected for each relation; the corresponding text is repeated as (13). The tree simplifies only in one respect: discourse markers are shown as words rather than as discourse marker specifications. Working through the tree, we see that this particular specification of the rhetorical structure from Figure 1 is a single sentence, composed of a single phrase (p1) followed by a complex of two phrases (p2 and then p3). The relation between p1 and p2+p3 is `concession`, lexicalised by 'but', and the feature CUE-STORE passes the information down, first to the internal node for the nucleus (as the lexicon tells us 'but' is realized there), then on to the terminal node for p2 where the marker will be realised, as p2 precedes p3. The choice of 'but' fixes the order of p1 and p2+p3, because its syntax requires that the nucleus be in second position. The `justify` relation between p2 and p3 is lexicalised by 'since', which is also realised on p2, its satellite (both orders for p2 and p3 will be generated).

(13) Elixir has no significant side-effects, **but since** the medicine is for you, never give Elixir to other patients.

Other variants will also be generated, for instance:

(14) Elixir has no significant side-effects. **However**, never give Elixir to other patients, **since** the medicine is for you.

(15) Elixir has no significant side-effects. **However**, **since** the medicine is for you, never give Elixir to other patients.

(16) Elixir has no significant side-effects.
However, the medicine is for you. **Consequently**, never give Elixir to other patients.

Two of the constraints which have been implemented to capture more stylistic preferences are:

- The root of the text structure should be a single unit of a level of sentence or higher, i.e. we do not want to generate isolated phrases or disconnected sequences of units.

[4] There is other syntactic information about each discourse marker, most critically what the syntactic forms are licensed with it, which is not shown here.

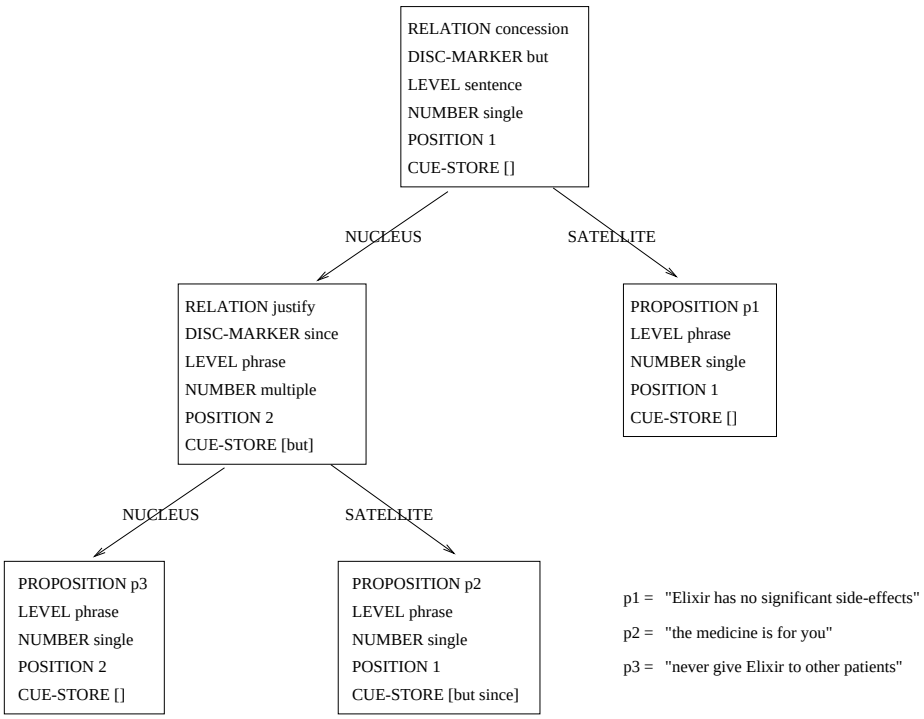

Figure 4: A formal text structure

- If a discourse marker is expressed within a complex unit, it should be attached to the simple unit that occurs first in the text.
  - Elixir has no significant side-effects. **However**, since the medicine is for you, never give Elixir to other patients.
  - #Elixir has no significant side-effects. Since the medicine is for you, never give Elixir to other patients, **however**.

  We are not claiming that texts which violate this constraint are ungrammatical, merely that they are more prone to ambiguity and, therefore, to misinterpretation.

4 Related work

Like Danlos (1998), we treat discourse markers as lexical realisations of certain clusters of features, and handle clauses, sentences and texts within a uniform framework. This work is also closely related to that of Webber and Joshi (1998) in that both use features to constrain the set of discourse markers which can appear in a particular context, and take into account both syntactic and rhetorical properties of discourse markers. It differs from their approach in that we specify the syntactic options within the feature system, while Webber and Joshi use LTAG trees which make explicit the syntactic configurations available to a given set of discourse markers.

Grote and Stede (1998) suggest a way of constructing a lexicon of discourse markers, which contains the appropriate features to allow a generation system to choose amongst them. Their proposal appears to be completely consistent with the present work. Many of the features they identify as important are the same as those which we use, and the additional information they

would include could easily be added to our representation (e.g. level of formality and polarity information).

5 Other interesting issues

There are a number of interesting issues which arise from this work, but which we have not yet had time to work out detailed solutions for. One is determining the contexts in which it is preferable to not mark a rhetorical relation explicitly, i.e. the issue of discourse marker *occurrence* mentioned at the start of the paper. Conversely, when is it appropriate to doubly mark a relation, as in example (17)? Another issue of interest to us is the conditions under which it is appropriate to generate highly under-specified discourse markers, such as ':' or 'and'?

(17) Elixir has no significant side-effects, **but, however,** since the medicine is for you, never give it to other patients.

We also want to extend our approach to correlative discourse markers, such as 'on the one hand...on the other hand' or 'if...then' (cf. Webber et. al's (1999) treatment of these markers in LTAG). Finally we want to treat a wider range of text-structures, including formatting information (e.g. indented lists) and other text units within sentences (e.g. Nunberg's text-clause level).

References

Robin Cohen. 1987. Analysing the structure of argumentative discourse. *Computational Linguistics*, 13(1–2):11–24.

Laurence Danlos. 1998. Linguistic ways for expressing a discourse relation. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Proceedings of the COLING-ACL'98 Workshop on Discourse Relations and Discourse Markers*, pages 50–53, Montreal.

Barbara Di Eugenio, Johanna D. Moore, and Massimo Paolucci. 1997. Learning Features that Predict Cue Usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL97)*, Madrid.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3).

Brigitte Grote and Manfred Stede. 1998. Discourse Marker Choice in Sentence Planning. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Ontario.

Alistair Knott and Chris Mellish. 1996. A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech*, 39(2–3):143–183.

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

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. Also available as USC/Information Sciences Institute Research Report RR-87-190.

Geoffrey Nunberg. 1990. *The Linguistics of Punctuation.* CSLI Lecture Notes, No. 18. Center for the Study of Language and Information, Stanford.

Rachel Reichman. 1981. *Plain-speaking: A theory and grammar of spontaneous discourse.* Ph.D. thesis, Dept. of Computer Science, Harvard University.

Dietmar Rösner and Manfred Stede. 1992. Customizing RST for the Automatic Production of Technical Manuals. In Robert Dale, Eduard Hovy, Dietmar Rösner, and Oliviero Stock, editors, *Aspects of Automated Natural Language Generation*, volume 587 of *Lecture Notes in AI*. Springer-Verlag, Heidelberg.

Donia Scott and Clarisse Sieckenius 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*, Cognitive Science Series. Academic Press.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Proceedings of the COLING-ACL'98 Workshop on Discourse Relations and Discourse Markers*, pages 86–92, Montreal.

Bonnie Webber, Alistair Knott, and Aravind Joshi. 1999. Multiple Discourse Connectives in a Lexicalized Grammar for Discourse. In Harry Bunt and Elias Thijsse, editors, *Proceedings of the Third International Workshop on Computational Semantics (IWCS-3)*, pages 309–325, Tilburg.