Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts*

Nadjet Bouayad-Agha
Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4AT, UK
nadjet@itri.brighton.ac.uk

Abstract

In order for a text planner to produce all the possible pragmatically congruent texts and only these, we distinguish between abstract and concrete rhetorical representations of a text. We discuss these representations and present our methodology for exploring the mappings from the underlying message to the actual surface discourse.

1 Introduction

We pose the following problem: what should be the input to a text planner in order to produce all the possible pragmatically equivalent texts and only these? This question is motivated by the aim to generate automatically patient information leaflets (PILs) in a variety of styles.[1] Indeed, these leaflets, often about the same medicine (or type of medicine), are produced in different house styles by various pharmaceutical companies. Our approach is constraint-based; that is, we aim to produce the set of all possible solutions, further reducing it with explicit constraints (Power, 2000). These constraints can be implemented into the system so that they always hold or they can be presented as input to the system, either as fine-grained choices or as general stylistic requirements (such as *formality*) which then trigger the relevant constraints.

We use Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) for representing our input as it allows a flexible operationalisation. For example, consider the following knowledge base:[2]

```
A: Mary cannot attend
B: Mary is ill
C: Mary has an exam
concession(sat:C,nuc:A)
evidence(sat:B,nuc:A)
```

Using a bottom-up text-plan construction algorithm such as Marcu's (1996), we are able to produce texts (1) and (2) below with their corresponding rhetorical structures in figure 1 (top and bottom respectively).

(1) Although Mary has an exam^C, since she is ill^B, she cannot attend^A.

(2) Since Mary is ill^B, although she has an exam^C, she cannot attend^A.

The two rhetorical structures in figure 1 are equivalent given the *nuclearity principle*, which states that "whenever two large text spans are connected through a rhetorical relation, that rhetorical relation holds also between the most important parts of the constituent spans" (Marcu, 1996). However, using only one structure as the input to the generator does not permit the production of both texts. In addition, some perfectly valid texts cannot be produced from those

*This work is undertaken within the ICONOCLAST project, which is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant L77102.

[1]The PILs corpus consists of more than 500 leaflets from about 50 different pharmaceutical companies, and adds up to 650,000 words. The corpus was made available electronically in HTML and SGML format (Bouayad-Agha, 2000).

[2]In the rhetorical assertions, `sat` corresponds to satellite and `nuc` to nucleus. To simplify, we represent the propositions as canned text.

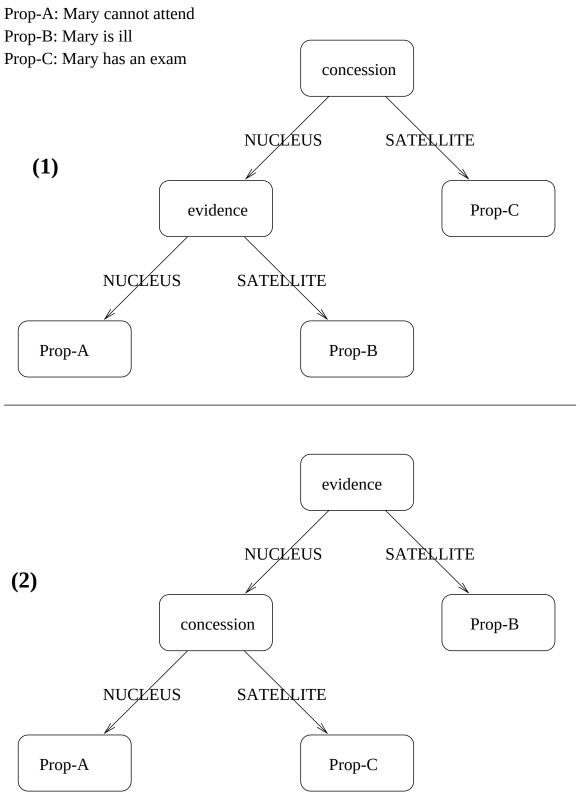

Figure 1: Rhetorical structures for text (1) and (2)

rhetorical structures. For example:

(3) Although Mary has an exam^C, she is ill^B. Therefore, she cannot attend^A.

Our work investigates the production of texts from rhetorical assertions. From a manual analysis of a few PILs, we observed discrepancies between the surface text and the rhetorical input. This suggests that influences other than the nuclearity principle play a role in planning acceptable text structures. Our current aim is to determine what these are.

These discrepancies between the surface text and rhetorical input lead us to posit a distinction between an *abstract* and a *concrete* rhetorical representation (ARR and CRR respectively) similar to the RAGS data models (Cahill et al., 1999).[3] The CRR is actually a hierarchical structure which is closer to the surface structure of the text; that is, it is the structure one would get when performing an RST analysis on the final linearised text. The ARR, on the other hand, is envisioned in two ways in the RAGS framework:

(I) It can be a set of semantically equivalent rhetorical structures, one of which is chosen to be the CRR. For example, the rhetorical structures in figure 1 can be considered to be semantically equivalent given the *nuclearity principle*.

(II) Alternatively, the ARR can be a more generic rhetorical representation, which is transformed into a CRR. These transfomations include, for example, repeating or aggregating nodes, or changing a relation from a more generic one to a more specific one. The only constraint on these transformations is that they are *meaning-preserving*.

In the work described here, we take the ARR to be the set of all relations needed to express the same meaning in different texts. This set can contain rhetorical assertions to be expressed disjointly or together, and it can also contain assertions whose arguments are embedded assertions, in which case we take the nuclearity principle not to apply at this level. The CRR, on the other hand can be a structure which results from the combination, segmentation and ordering of these assertions using syntactic, semantic and aggregation rules, the simplest of which is the nuclearity principle. This is illustrated in section 2, with real examples from the PILs corpus.

The distinction between abstract and concrete rhetorical representations has already been used by Delin et al. (1996) to distinguish between a language-independent semantic representation and its actual rhetorical realisations in different languages. Thus, the problem of finding paraphrases in the same language is similar to that of producing texts in different languages.

The aim in studying these two representations is to find the constraints (or transformations) for deriving CRRs from ARRs and

[3]RAGS' aim is to provide a Reference Architecture for Generation Systems with the description of generic data models.

to formalise them into text planning rules, so that a text planner is able to produce *all* and *only* the pragmatically congruent texts that express the ARR. This investigation is being carried out mainly through a corpus analysis which is described in section 3.

2 Abstract and concrete rhetorical representations

Examination of real texts quickly reveals that their surface realisation leaves out some relations that can be inferred by the reader. For example, consider (4a) and (5a) below:[4]

(4a) If you are pregnant or breast feeding^A, do not take Elixir tablets without consulting your doctor first^B, as the safety of Elixir tablets in pregnancy and breast feeding is not known^C.

(5a) If you have difficulty in swallowing tablets^A, consult your doctor^B, as he/she may wish to change your medicine^C.

Both have the same surface rhetorical structure (i.e., CRR) which can be represented as in figure 2. This rhetorical structure, given the nuclearity principle, amounts to the following two rhetorical assertions:

```
condition(sat:A,nuc:B)
evidence(sat:C,nuc:B)
```

However, (5a) has an extra relation which is `condition(sat:A,nuc:C)`, as (5b) in contrast with (4b) illustrates (both texts are made-up paraphrases). This assertion `condition(sat:A,nuc:C)` can be used interchangeably with assertion `condition(sat:A,nuc:B)` to convey the message and both can be retrieved from the text by the reader.

[4]All the examples in this paper are taken from the PILs corpus, with no modifications apart from the alteration of the product names for legal reasons, unless stated otherwise.

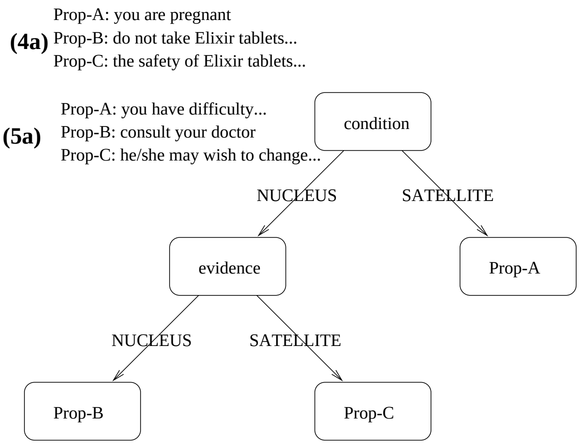

Figure 2: Surface Rhetorical structure for texts (4a) and (5a)

(4b) # If you are pregnant or breast feeding^A, the safety of Elixir tablets in pregnancy and breast feeding is not known^C. Therefore, do not take Elixir tablets without consulting your doctor first^B.

(5b) If you have difficulty in swallowing tablets^A, your doctor may wish to change your medicine^C. Therefore, you should consult him/her^B.

Although RST was not designed to handle hypothetical contexts[5], this kind of mismatch between surface segmentation and ARR is widespread in the PILs corpus, not only with conditionals, but also for other hypothetical situations such as CIRCUMSTANCE or PURPOSE. Therefore, a text planner needs to take this phenomenon into account in order to produce such texts. An intuitive ARR to deal with example (5a) would be the structure in figure 2, where the nuclearity principle is abolished at this abstract level. On the other hand, the ARR for (4a) can be the set of assertions between propositions. Given the same ARR for (5a), (5b) is structurally incompatible with it since A and C are grouped together in a sentence.

This structural incompatibility between rhetorical and text structures is discussed in more detail in a recent paper by Bouayad et

[5]Discourse Representation Theory (DRT) provides complex semantic devices to account for this phenomenon (e.g., Franck and Kamp (1997)).

al (2000). In this respect, note that structural incompatibility may lead to perfectly acceptable paradoxical surface semantic structures, as the examples (6) and (7) below show. In (6), B and A are grouped together and a purpose relation is expressed between them. However, given the rhetorical relations definitions in (Mann and Thompson, 1987), B is in fact the purpose of C which is in a kind of contrast with A. In (7), there is a concession relation between A and B (at least) on the surface, since the adverbial *However* is appended before B. But the CONCESSION semantically holds between A and C, which is in an evidence relation with B. This seems to be licensed by the tri-partite concession relation as described in (Grote et al., 1997).

(6) (If you get wheezy) don't use Elixihaler^A to stop your wheeze^B. You should use your "reliever" puffer^C, just like you did before.

(7) (The label will tell you how much to take and how often.) Usually this will be twice a day, a dose in the morning and one in the evening^A. [..] However, all patients are different^B and your doctor may have prescribed more or less for you^C.

Other phenomena which we are able to capture by distinguishing between ARR and CRR is aggregation, as text (8) below illustrates:

(8) It is estimated that up to half a million children catch head lice each year^A. Also, as with the common cold, anyone can catch head lice^B. So if someone in your family catches head lice^C, they are not alone^D, and it is certainly nothing to be ashamed about^E.

where the ARR could look like the following:

```
evidence(sat:A,nuc:D)
evidence(sat:B,nuc:E)
condition(sat:C,nuc:D)
condition(sat:C,nuc:E)
```

whereas in the CRR, A and B are joined together (marker *Also*), providing evidence (marker *So*) for D together with E under a common condition C.

3 Methodology

In order to study more thoroughly the mechanisms for deriving surface representations from abstract ones, we acquired a corpus of multi-clause texts from the PILs corpus, with distinct ARR and CRR information, that is, texts whose CRR is not obtained by simply combining the rhetorical assertions of the ARR using the nuclearity principle. The ARR is deduced from paraphrases. These can be obtained manually by the annotator, with the help of the implemented text planner described at the end of this section (stage VI) and by finding (if possible) real paraphrases across different leaflets. The CRR is obtained from surface clues such as syntax and text segmentation. In this section, we present our methodolgy for studying these representations. This methodology involves six stages presented below. To date, the first three preliminary stages have already been completed whereas the last three core stages are currently being carried out.

(I) Preparing the data. All the paragraphs from the PILs corpus which contain explicit discourse markers were extracted. The discourse markers used were the most commonly found markers (preference given to the argumentative ones) in a set of 20 manually analysed PILs, namely: *if, even if, in order to, to (*PURPOSE*), by (*ENABLEMENT*), so, because, since, but, however, although, as, otherwise, unless.* After removing the duplicated paragraphs using the method described in (Bouayad-Agha and Kilgarriff, 1999) and the two-clauses paragraphs, there were around 3000 paragraphs left.

(II) Finding the text fragments for the study. In total around 500 out of the 3000 paragraphs present discrepancies between the ARR and the CRR. Three criteria were used for extracting those texts:

Textual: The text structure, that is the division of the text in text-clauses, text-sentences and paragraphs (following Nunberg's (1990) text grammar) is not compatible with its abstract rhetorical structure. In other words, there are some groupings in the text structure that are not present in the rhetorical structure. This principle is enunciated in (Power, 2000; Bouayad-Agha et al., 2000). Thus, (5b), (6) and (8) would be picked up using this criterion.

Syntactic: The syntactic structure is incompatible with the abstract rhetorical structure. For example, in (9) below, B is syntactically subordinated to A and C is coordinated to the complex clause A-B whereas C is rhetorically in contrast with B. Similarly, in (7), the adverbial *However* is attached before B which is not the nucleus of the CONCESSION whereas in fact, it should be attached to C. This criterion takes into account the sometimes incompatible two-level discourse representations, intentional and informational, noted by some researchers (Moore and Pollack, 1992; Moser and Moore, 1996).

(9) Elixir can sometimes cause allergic reactions^A [...] which are usually mild^B, but very rarely allergic reactions can cause difficulty in breathing, fainting and swelling of the face and throat^C.

Semantic: The ARR is either a structure which relaxes the nuclearity principle (5a) or a disjunction of assertions. Text (10a) is an example of the latter, where the concessive condition C can attach either to B or A, with no change of meaning, as its paraphrase (10b), which can actually be found in the PILs corpus as well, shows.

(10a) (These tablets/sachets are for you.) Never give them to someone else.^{A} This medicine may harm them^B, even if their symptoms are the same as yours^C.

(10b) (These tablets/sachets are for you.) Never give them to someone else^A even if their symptoms are the same as yours^C. This medicine may harm them^B.

(III) Surface Annotation. This task involves annotating the discourse markers, text segments and elementary discourse units (EDUs) (Marcu, 1996) together with their syntactic type as SGML tags on the text. A simple Document Type Definition (DTD) and an annotation procedure were devised for this purpose. The annotation was initially done automatically, the EDUs being annotated as enclosed within the text segments, within which they were tagged according to the delimitation of commas.[6] All the EDUs as well as markers have a unique ID within each record. An example of this surface annotation corresponding to text (5b) is given below.

```
<record id=1000 filename="example.sgml">
<txtseg type="paragraph">
<txtseg type="text-sentence">
<marker id=0>If </marker>
<edu id=0 type="sentence">
you have difficulty in swallowing tablets,
</edu>
<edu id=1 type="sentence">
your doctor may wish to change your medicine.
</edu>
</txtseg>
<txtseg type="text-sentence">
<marker id=1>Therefore,</marker>
<edu id=0 type="sentence">
you should consult him/her.
</edu>
</txtseg>
</txtseg>
</record>
```

(IV) CRR Annotation. Every paragraph corresponds to an ordered surface rhetorical structure tree with marker information on the branch nodes. For instance, for text (5b), the CRR is as follows, with the numbers on the PROP elements corresponding to the EDUs' IDs and on the MARKER elements to the markers' IDs in the surface annotation.

[6] Commas are not considered to be a reliable indicator of text structure: for one thing, a comma can usually be omitted as it often corresponds to syntactic boundaries.

```
<crr id=1000>
<relation name="evidence">
<satellite>
<relation name="condition">
<marker id="0">
<satellite><prop id="0"></satellite>
<nucleus><prop id="1"></nucleus>
</relation>
</satellite>
<nucleus>
<marker id="1">
<prop id="2">
</nucleus>
</relation>
</crr>
```

(V) ARR Annotation. The abstract representation is expressed by a conjunction and disjunction of rhetorical assertions, whose elements are either propositions or rhetorical assertions. The ARR for (5b) is as follows:

```
<arr id=1000>
<relation name="condition">
<satellite><prop id="0"></satellite>
<nucleus>
<relation name="evidence">
<satellite><prop id="1"></satellite>
<nucleus><prop id="2"></nucleus>
</relation>
</nucleus>
</relation>
</arr>
```

(VI) ARR/CRR Analysis. Given this annotation, we want to devise operations of transformation from the ARR to the CRR in a similar way to Marcu's (2000) transfer rules for translating a Japanese discourse configuration into an English one. Since the corpus is limited in size and only provides acceptable data, we have also implemented in Prolog a small system which, given a set of rhetorical assertions in which the propositions are represented as canned text, produces a variety of output texts by building all the possible CRRs, regardless of nuclearity. The only constraint is that the relation that is expressed between two spans holds *at least* between one proposition in each of those spans. This is equivalent to producing all the possible ordered binary trees given a number of leaves. This method can at present only efficiently work for a limited number of clauses, since, say, for 5 clauses, there are 168 solutions whereas for 6 clauses there are 30240 solutions.[7] Currently, we are generating all the possible twelve CRRs given 3 propositions and a set of rhetorical assertions holding between them. This way, we are able to test each ARR in the corpus by producing all its possible surface realisations, and devise constraints on the text planner and the abstract representation in order to produce only meaning-preserving texts.[8] For example, from the CRR and ARR annotations of text (7) reproduced below, we may assume that this construction is always possible when these two relations are involved in this particular configuration. We can test this hypothesis by generating a number of examples. Example (11) shows, however, that this surface configuration can lead to an ambiguity between A conceding over B (*John did not come although there was lots of sweets...*) or A conceding over C (*Mary did not enjoy herself although there was lots of sweets*). We concluded that this configuration is only possible if no conceding relation between A and B can be inferred by the reader. Another factor which has been found to constrain some incompatible structures is linear order.

(7) (The label will tell you how much to take and how often.) Usually this will be twice a day, a dose in the morning and one in the evening^{A}. [..] However, all patients are different^{B} and your doctor may have prescribed more or less for you^{C}.

(11) There was lots of sweets and chocolates at the party^{A}. However, John did not come^{B} and (therefore,) Mary did not enjoy herself^{C}.

4 Conclusion

The distinction between *abstract* (ARR) and *concrete* (CRR) rhetorical representations is not an absolute requirement for generating a valid text plan; after all, most text planners do pretty well without them. However, we

[7] This is calculated with the following formula, given that there are N+1 propositions: $N!(1/(N+1))(C_N^{2N})$ (Sedgewick and Flajolet, 1996).

[8] The assignment of discourse markers is performed as described in (Power et al., 1999).

argue here that this distinction is required in any context where the goal is to produce a range of variants of a given message to be delivered. This is the case in our application, where leaflets describing the same or similar medicines need to be generated in different styles for different manufacturers. But the issue is not particular to this application. Indeed, it applies to most situations where texts need to be tailored to specific audiences, and to any system that aims to generate paraphrases. We believe that this approach also provides a way for properly capturing, and therefore generating, the divergent rhetorical structures of multilingual texts expressing the same content.

Acknowledgements

I would like to thank Donia Scott, Kathy McKeown and Richard Power for their valuable comments on earlier versions of this paper.

References

N. Bouayad-Agha and A. Kilgarriff. 1999. Duplication in corpora. In *Proceedings of the Second CLUK Colloquium*, Colchester, Essex.

N. Bouayad-Agha, R. Power, and D. Scott. 2000. Can text structure be incompatible with rhetorical structure? In *Proceedings of the International Natural Language Generation Conference*, pages 194–200, Mitzpe Ramon, Israel, 12–16 June.

N. Bouayad-Agha. 2000. Layout annotation in a corpus of patient information leaflets. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Athens.

L. Cahill, C. Doran, R. Evans, C. Mellish, D. Paiva, M. Reape, D. Scott, and N. Tipper. 1999. Towards a reference architecture for natural language generation systems. Technical Report ITRI-99-14, Information Technology Research Institute, March.

J. Delin, D.R Scott, and H. Hartley. 1996. Pragmatic congruence through language-specific mappings from semantics to syntax. Technical Report ITRI-96-01, Information Technology Research Institute, University of Brighton. A shorter version of this paper appears in Proc. 16th COLING, Copenhagen, August 1996.

A. Frank and H. Kamp. 1997. On context dependence in modal constructions. In *Proceeding of SALT 7*, Stanford University, March 21-23.

B. Grote, N. Lenke, and M. Stede. 1997. Ma(r)king concessions in english and german. *Discourse Processes*, 24(1):87–117.

W.C. Mann and S.A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, Information Sciences Institute, University of Southern California.

D. Marcu, L. Carlson, and M. Watanabe. 2000. The automatic translation of discourse structures. In *Proceedings of the North American Chapter of the Association for Comptuational Linguistics (NAACL'2000)*, Seattle, Washington.

D. Marcu. 1996. Building up rhetorical structure trees. In *The Proceedings of the Thirteenth National Conference on Artificial Intelligence*, volume 2, pages 1069–1074, Portland, Oregon, August. AAAI.

J.D. Moore and M.E. Pollack. 1992. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

M. Moser and J.D. Moore. 1996. Towards a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

G. Nunberg. 1990. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. CSLI Publications, Stanford, CA.

R. Power, C. Doran, and D. Scott. 1999. Generating embedded discourse markers from rhetorical structure. In *The Proceedings of the European Workshop on Natural Language Generation*, pages 30–38, Toulouse. Also available as a technical report ITRI-99-15 at http://www.itri.brighton.ac.uk/techreports/.

R. Power. 2000. Planning texts by constraint satisfaction. In *The 18th International Conference in Computational Linguistics (COLING)*, Saarbrücken.

R. Sedgewick and P. Flajolet. 1996. *An Introduction to the Analysis of Algorithms*. Addison-Wesley Publishing Company.