Multiple Discourse Marker Occurrence: Creating Hierarchies for Natural Language Generation

Sarah Louise Oates*
University of Brighton.
Sarah.Oates@itri.brighton.ac.uk

Abstract

Most studies on discourse markers implicitly assume that only one marker or discourse relation will occur in a sentence. In reality, more than one relation may hold between text spans and may be cued by multiple discourse markers. We describe here a method for hierarchically organising discourse markers. The hierarchies are intended for use by a generation system to enable the selection and placement of more than one marker in a single text span.

1 Introduction

The majority of studies on discourse markers implicitly assume that only one marker or discourse relation will occur in a sentence or that the presence of multiple markers will not affect the choice and placement of others. However, in reality, more than one relation may hold between text spans which may be cued by multiple markers. The available rules describing the occurrence, choice and placement of a given marker do not account for multiple marker occurrence (Grote et al., 1995; Webber and Joshi, 1998; Power et al., 1999, e.g.,). We have found that the choice and placement of discourse markers is greatly affected, not only by the presence and number of other markers, but also by the style of the text and the strength of other markers in the text span. We describe here a method for hierarchically organising discourse markers which takes account of these factors. The hierarchies are intended for use by a generation system to enable the selection and placement of multiple markers.

* The author would like to thank the Engineering and Physical Sciences Research Council for funding

2 Defining Discourse Markers

Although precise definitions of discourse markers differ between studies, it is generally accepted that their role is to signal how one proposition should be interpreted given the other(s) in the discourse (Millis et al., 1995; Moore and Pollack, 1992). Most researchers in this field also agree that the relation between these propositions may exist regardless of whether a discourse marker is used (Scott and de Souza, 1990; Knott, 1995): a discourse marker is simply an explicit signal of a specific relation between two or more propositions. The non-occurrence of a marker does not mean that a discourse relation is absent:

(1) *no marker, 1 relation:* The museum does not intend to sponsor a particular aspect of modern art; it intends to make a report to the public by offering material for study and comparison.

By the same token, the presence of more than one discourse marker does not always signal a multitude of relations:

(2) *2 markers, 1 relation:* The museum does not intend to sponsor a particular aspect of modern art, **but rather** to make a report to the public by offering material for study and comparison. (BNC)[1]

Previous studies have accounted for a wide range of phenomena, from choosing between similar discourse markers (Fraser, 1998; Sanders et al., 1992) to abstracting away from discourse markers and using syntax to signal underlying discourse relations (Delin et al., 1996). However, the issue of multiple markers, like those in the example above, is only now beginning

[1] British National Corpus (Leech et al., 1994)

to be addressed. Recent work in computational linguistics has provided possible solutions for the use of correlative markers (Webber and Joshi, 1998) and embedded clauses (Power et al., 1999). However, these solutions are incomplete and further research is needed if we are to account for all examples of multiple discourse markers.

3 Multiple Markers

The present project focuses on *all* cases of *multiple discourse markers*, in other words, all cases where more than one marker occurs within two spans of text which are expressed either (a) within the same *text sentence* (Nunberg, 1990) covering one or more discourse relations (e.g., examples 3 and 4);

(3) **Having said that, if** you weigh only 60 kg (132lb) **and yet** still manage to sit your 90 kg (198lb) opponent down with a solid thump to his mid-section, **then** the refereeing panel may well applaud your fervour with a full point. (BNC)

(4) **Since** the question turns on the meaning of the word "appropriate" in section 1(1) of the Act of 1968, the problem is **therefore** one of statutory interpretation. (BNC)

or (b) in different text sentences but covering only one relation, the so-called *correlative markers* (Quirk et al., 1985) (e.g., example 5):

(5) The job of being an Acorn Project leader is an unenviable one. **For a start**, they don't get paid, though they do receive a petrol allowance; **for another thing**, it's a bit like being in a group of unruly children for the week... (BNC)

The work described here focuses solely on *multiple discourse markers cueing a single relation*, paying attention, when possible, to embedded discourse relations and their markers.

4 Single Relations — Multiple Markers

Preliminary tests using the British National Corpus (BNC) and Knott's (1995) taxonomy of discourse markers suggested that the order of multiple markers cueing a single relation is affected by their position in the taxonomy; those higher in the taxonomy always precede those lower in the taxonomy (see figure 1 and examples 6-7);

(6) This blood-line was particularly helpful to the early breeders because the line was in-bred, his parents being brother and sister of excellent breeding **and so consequently** true to type. (BNC)

(7) The difficulty is that the sites which have been extensively excavated, **and so** produced the largest quantities of pottery, such a Corbridge and Newstead, are multi-period, and the stratification of the excavations early in the century, **consequently** suspect. (BNC)

However, since Knott's taxonomy only allows us to view hierarchies of markers of a single relation, improvements were necessary in order to account for multiple markers. Using the BNC, a list of at least 350 English discourse markers and Mann and Thompson's (1988) original 23 rhetorical relations, we created a database on the number and type of relations each marker can cue (see figure 2). From this a hierarchy was built, similar to Knott's (1995), but benefiting from a wider range of markers and allowing more than one relation to be expressed at a time, thus reducing the redundancy present in Knott's taxonomy. Furthermore, in contrast to Knott's study in which examples were fabricated, all examples of discourse marker usage in our database are taken from the British National Corpus (BNC). Thus, all of our examples are taken from real, natural texts and are, therefore, representative of discourse marker occurrence in natural language.

5 Constructing the Hierarchy

Our hierarchies are constructed on the assumption that (a) some discourse markers may be used to cue more than one relation and (b) when more than one marker is needed, the number of relations a marker can cue will affect the choice and position of that marker. In our hierarchy, those discourse markers which can cue many relations appear at the top and those marking only a single relation occur at the bottom. Markers may also have additional constraints on their usage depending upon the text style, other relations being marked simultaneously and the content of the related propositions.

6 Strong & Weak Markers

Figure 3 is an example of our hierarchy for the family of contrastive relations. Here we see that 'but' can mark four discourse relations

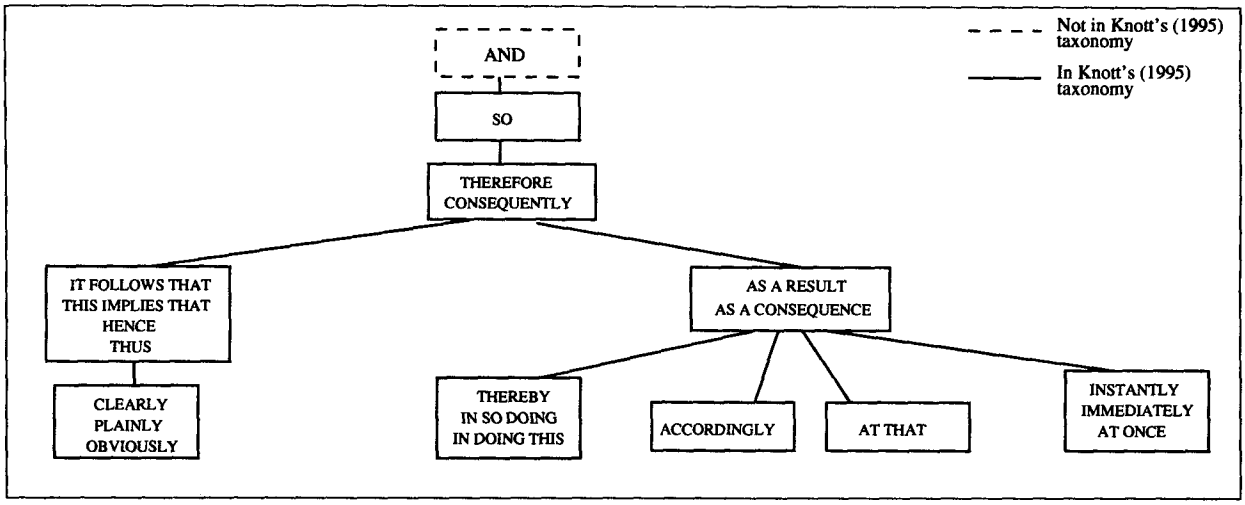

Figure 1: Example of Knott's (1995) Taxonomy

DISCOURSE MARKER	CATEGORY	DISCOURSE RELATION	SAT/ NUC*	EXAMPLE IN USE
so	subordinator	vol-cause	(n)	He had no chance of winning SO he pretended he wasn't trying (BNC)
		non-vol-cause	(n)	While deciding to stay as independent as possible, I contacted ACET who I knew provided practical care at home. I had previously spent about 2 years asking local services and friends for help and not having it happen, SO my flat had become pretty run down. (BNC)
		non-vol-result	(s)	While wanting to dismiss the stereotyping and silly superstition, the snag remains that within all the ballyhoo there are elements of truth. SO instead of being outraged, one is left with a resigned smirk. (BNC)
		enablement	(n)	Loosen the cord SO you can remove the curtains easily. (BNC)
		evaluation	(s)	Nor is this feeling only provoked by the sight or the thought of art, he wrote. I also experienced it when I signed the marriage register as well as when I saw the pig slaughtered...a feeling of the heart leaping and the blood pumping.....SO, wrote Harsnet, there is continuity as well as discontinuity. (BNC)
		justify	(n)	If you went on strike they didn't pay you off. You got sacked and you just didn't get any money. So people had no other option but to work. (BNC)
		sequence	multi-nuclear	.....that's what I guessed so I said "no" , I said they're fine, SO she said "oh, I'm ever so sorry". I said "don't be". (BNC)
		purpose	(s)	He'll remind her SO she'll remember. (BNC)

*SAT/NUC = The text span upon which the discourse marker occurs - SAT(satellite), NUC(nucleus).

Figure 2: Extract from Database of Discourse Markers & Relations

(contrast, antithesis, concession and exception) without constraint. When discourse markers can be used for a large number of relations, we refer to them as 'weak' markers since there is only a weak correlation between the marker and the relation being signalled. In contrast, when a discourse marker can only cue a single relation, we refer to it as a 'strong' marker, since there is a strong correlation between the relation and the explicit lexical cue. In the hierarchy 'notwithstanding that' is a highly constrained, strong discourse marker since it can only mark one relation (concession) and occur only when the text is formal, legal or both.

Our tests on the BNC show that the choice and placement of a marker will be affected by its strength or weakness; the weakest markers always precede the stronger ones. We find that this rule not only applies to single relations cued by multiple markers:

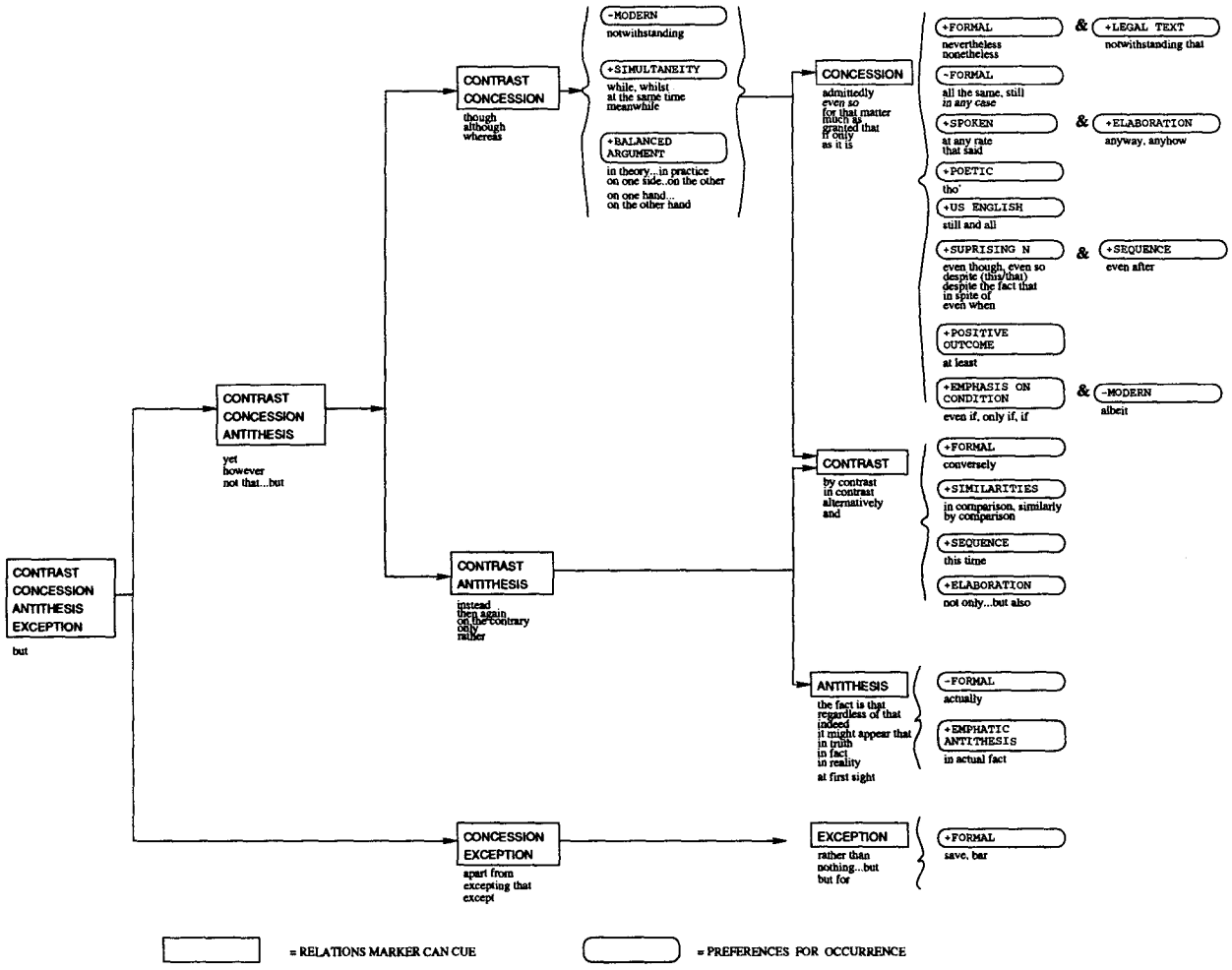

Figure 3: Hierarchy of Contrastive Family of Relations

(8) (a) The pores in the skin are a classic example: they can not become perceptible to us by themselves, **but yet** (b) their presence in the skin can be deduced from sweat. (BNC)

but also, to a certain extent, to embedded relations cued by two markers in the same text span. In the following example, we have two relations and two markers of contrast. The superordinate relation, marked by 'however', holds between proposition (a) and propositions (b) and (c), whilst the subordinate relation, marked by 'whereas', holds between propositions (b) and (c);

(9) Indeed , (a) so strong have the differential views on advantageous locations become that one recent assessment of the total stock of foreign capital in developing countries suggests that it is less today than it was in 1900. **However, whereas** (b) the G-5 countries now account for 75 per cent of the world's FDI flow, (c) their position as the five major exporters is a much less concentrated 45 per cent.(BNC)

In both cases, the weakest marker precedes the stronger marker and neither could be reversed and remain grammatical. Thus, working through the hierarchy from the weakest to the strongest markers, a generation system can determine which discourse marker should occur in a particular position on the basis that the weakest markers always precede the stronger ones.

Decisions are based on the relation(s) to be marked, any other relation(s) already present, the style of the text, the content of the text spans, and the strength or weakness of other discourse markers present.

7 Conclusions

Thus far, we have developed hierarchies for the family of contrastive relations in English and French and the family of causal relations in English. Ultimately, we intend to establish a complete hierarchy of all the markers of discourse relations; this will not only allow us to choose between different markers, regardless of whether one or more are used, but will also help to determine their order when multiple markers are necessary. In the final version of the hierarchy, we intend to provide the generation system with statistical information on the likelihood of one marker following another. Such information will take account of the fact that certain markers tend to occur together more often than others. These statistics are currently being derived from tests on the International Corpus of English (Nelson, 1995).

References

J. Delin, D. R. Scott, and A. Hartley. 1996. Language Specific Mappings from Semantics to Syntax. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 292–297. COLING96.

B. Fraser. 1998. Contrastive Discourse Markers in English. In A. H. Jucker and Y. Ziv, editors, *Discourse Markers: Descriptions and Theory*, pages 301–326. John Benjamins, Amsterdam.

B. Grote, N. Lenke, and M. Stede. 1995. Ma(r)king Concessions in English and German. In *Proceedings of the 5th European Workshop on Natural Language Generation*, pages 11–32, Leiden, May. Leiden University.

A. Knott. 1995. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* Ph.D. thesis, University of Edinburgh.

G. Leech, R. Garside, and M. Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 622–628, Kyoto, Japan.

W. Mann and S. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organisation. *Text*, 8:243–281.

K. K. Millis, J. M. Golding, and G. Barker. 1995. Causal Connectives Increase Inference Generation. *Discourse Processes*, 20(1):29–50.

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.

G Nelson. 1995. The International Corpus of English: Markup & Transcription. In G. N. Leech, G. Myers, and J. A. Thomas, editors, *Spoken English on Computer: Transcription, mark-up and application*, pages 220–223. Longman, London.

G. Nunberg. 1990. The Linguistics of Punctuation. CSLI Lecture Notes, no.18. Center for the study of Language and information, Stanford.

R. Power, D. Scott, and C. Doran. 1999. Generating Embedded Discourse Markers from Rhetorical Structure. In *Proceedings from the European Workshop on Natural Language Generation EWNLG'99*, pages 30–38, Toulouse.

R. Quirk, S. Greenbaum, S. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, London.

T. Sanders, W. Spooren, and L. Noordman. 1992. Toward a Taxonomy of Coherence Relations. *Discourse Processes*, 15:1–35.

D.R. Scott and C.S. 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, London.

B. L. Webber and A. K. Joshi. 1998. Anchoring a Lexicalised Tree-Adjoining Grammar for Discourse. In M. Stede, L. Warner, and E. Hovy, editors, *Discourse Relations and Discourse Markers. Proceedings from the Workshop. COLING-ACL'98*, pages 86–92, Montreal.