Pages 13–14 in *Working Notes for the AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, Asilomar, November 1991.

The Role of Punctuation in Discourse Structure

Robert Dale
Centre for Cognitive Science and Department of Artificial Intelligence
University of Edinburgh
Edinburgh EH8 9LW
Scotland
Email: R.Dale@edinburgh.ac.uk

Introduction

This paper takes the view that there are three distinct but interdependent systems available in written natural language texts for indicating the structure of discourse:

- lexical markers, including cue words and phrases;
- punctuation markers, including commas, colons, semi-colons, dashes and parentheses; and
- graphical markers, including the use of paragraph breaks and itemized or enumerated lists.

The first of these is clearly uncontroversial; it is only very recently, however, that researchers have begun to take the others seriously. As noted by Nunberg [1990], a pervasive bias towards taking the primary object of linguistic study to be the spoken form of language has led linguists to almost completely ignore phenomena that belong to the second and third of these categories. Nonetheles, it is clear that punctuation and graphical markers do play an important role in indicating structural relations in written discourse; if we are in the business of building systems for the automatic generation or analysis of written documents, we have to ensure that these systems incorporate an adequate model of these aspects of text.

Some recent work has broken the mould; Nunberg [1990] presents the beginnings of a theory of the linguistics of punctuation, and Hovy and Arens [1991] have experimented with the inclusion of text formatting commands in the ouptut of a text generation system. This paper raises some questions that arise in the context of integrating a theory of punctuation into a model of text structure. The basic claim is that, by ignoring these structure marking devices, our systems and theories have omitted an important element of meaning in written texts.

Some Data

Consider the natural interpretations of each of the following examples (borrowed from Nunberg [1990]):

(1) He reported the decision: we were forbidden to speak with the chairman directly.

(2) He reported the decision; we were forbidden to speak with the chairman directly.

(3) He reported the decision—we were forbidden to speak with the chairman directly.

We can characterise the differences here in terms of a theory like Rhetorical Structure Theory; examples (1) and (2) would exhibit instances of ELABORATION and CAUSE respectively, and example (3) could express either of these rhetorical relations. This makes it clear that punctuation markers at least play some role in the interpretation of particular rhetorical relations. The mapping is not straightforwardly one-to-one, however: at the very least, different relations may be realized by means of the same punctuation markers.

The Nature of the Underdetermination

From the examples above, we see that the particular rhetorical relations that reside in a text are underdetermined by the punctuation marks that are used. Note that the situation here is not all that different in kind to that which happens in the case of lexical markers, since the same cue words can be used for a number of different relations. Suppose, for example, we adopt a taxonomy of rhetorical relations like that proposed in Hovy [1990]: we can ask whether the use of a particular lexical marker can be rooted at some node in the taxonomic tree, with all the relations which are descendents of this node being able to employ that particular lexical marker as a clue to the relation being expressed (and correspondingly, any use of this marker being ambiguous with respect to which of the relations in this subtree it serves to realize). Given such an approach, we can ask whether it is possible to determine a similar assignment of punctuation markers to classes of relations. Going in the other direction, the distribution of punctuation markers may tell us something about how the space of rhetorical relations should be structured.

Closer examination suggests that at least some of the punctuation markers are so lacking in specificity that they tell us not very much at all about the particular relation they realize: consider the following example (again borrowed from Nunberg [1990]).

(4) Some people found the book fatuous; John considered it a paramount example of post-modern criticism.

Depending on the context, we can view the relation being expressed here as one of ELABORATION or one of ANTITHESIS. This suggests that, if we are to root the use of specific punctuation markers anywhere in a taxonomy of relations, the range of possible relations they may be realizing is so broad that the markers reside at the least specific nodes of the taxonomy. One possibility is that the punctuation markers correspond purely to *syntactic* structural relations in discourse, something like Grosz and Sidner's [1987] Dominance (DOM) and Satisfaction Precedes (SP) relations, but say nothing about the particular *semantics* of the relations.

Discourse Structure and Grain Size

Notice that the writer is often free to use structure indicators from any of the three systems. So, for example, in the context of generating recipes (cf [Dale 1990]), the writer can choose between a wide range of alternatives which express the same hierarchical and rhetorical relationships. Quite apart from the use of explicit lexical markers, for example, she can choose to express express related operations by means of semi-colon conjoined clauses, as in

(5) Soak the beans; drain and rinse them.

or by using more graphical devices, as in:

(6) 1. Soak the beans.
2. Then:
(a) Drain them.
(b) Rinse them.

Given the simple domain, this last example is somewhat artificial, but the general idea should be clear; mechanisms of this kind are often used in instruction manuals (again, see Hovy and Arens [1991] in this connection).

There are questions that then arise, of course, as to the constraints that hold between the three sets of markers; why should the author choose one rather than another? One criterion seems to be the size of the elements to be related: in general, the elements related by colons and semi-colons cannot extend beyond the sentence in which the marker appears. The use of punctuation is constrained in a non-trivial manner by the grain-size of the elements that are to be related.

Note that punctuation plays a role internal to major clauses as well as between them; following Scott and Souza's [1990] heuristics for communicative efficiency in text generation, we might fold information into minor clauses within a sentence, preferring (8) over (7):

(7) Knox is enroute to Sasebo. It is C4.

(8) Knox, which is C4, is enroute to Sasebo.

Again, the role of the punctuation marks is crucial here. Note, however, that theories of discourse structure generally restrict themselves to relating elements that are at least major clauses, and have little or nothing to say about sentence internal phenomena; but the alternative realisations available in (5) and (6), and the need to view the non-restrictive relative clause in (8) as partaking in some rhetorical relation, suggest that we have some motivation for looking for a theory of discourse structure that operates both above and below the level of the sentence.

Towards this end, we can view the use of these punctuation marks as simply one possible realization of hierarchical structural relations that could be realized in other ways. The choice of this system of indicators is then tightly knit with decisions such as the chunking of information into sections, paragraphs and sentences, and the folding in and restructuring of material into a text using adjuncts both above and below the level of the sentence. We can argue that discourse structure below the sentence is a reality, made obvious when we look at written texts; a prior emphasis on spoken texts has been deceptive in this respect.

This also raises questions about constraints on anaphora: can we describe the anaphoric relations that hold within complex, multi-clausal sentences using similar mechanisms to those that have been proposed as holding between larger discourse segments? The fact that, for example, anaphoric reference to elements within a parenthetical element (such as this phrase) is not possible from outside the parenthetical element might be explained in this way. This phenomenon, of course, is additional justification for considering discourse structure to be relevant below the level of the sentence.

References

R Dale [1990] Generating Recipes: An Overview of Epicure. In R Dale, C Mellish and M Zock (eds), *Current Research in Natural Language Generation.* Academic Press.

B Grosz and C Sidner [1987] Attention, Intentions and the Structure of Discourse. *Computational Linguistics.*

E Hovy [1990] Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. In *Proceedings of the Fifth International Natural Language Generation Workshop*, Dawson, Pennsylvania.

E Hovy and Y Arens [1991] Automatic Generation of Formatted Text. In *Proceedings of AAAI–91.*

G Nunberg [1990] *The Linguistics of Punctuation.* CSLI Lecture Notes No. 18, University of Chicago Press.

D Scott and C S de Souza [1990] Getting the Message Across in RST-based Text Generation. In R Dale, C Mellish and M Zock (eds), *Current Research in Natural Language Generation.* Academic Press.