Pages 110–120 in *Proceedings of a Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI*, Technical University of Berlin, 9th–11th October 1991.

Exploring the Role of Punctuation in the Signalling of 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

1 Introduction

This paper reports on the beginnings of some work whose aim is to investigate the role of punctuation in signalling discourse structure. This is part of a larger enterprise motivated by the view that natural language processing systems, whether generating or understanding language, should take account of the physical manifestation of language: in the context of written language, this means taking account of the constraints imposed and opportunities offered by the real estate of the pages (or whatever other surfaces) on which the words appear. In this context, punctuation marks occupy an interesting middle ground between overtly linguistic phenomena like cue words and phrases, and overtly layout-oriented phenomena such as indentation and spacing. Ultimately, we need some means of representation that can accommodate both these dimensions of language use. This paper suggests that a study of the uses of punctuation can be a useful first step towards identifying what the elements of such a representation language should be.

2 Background

Language is always EMBODIED or SITUATED in some medium. This observation is an obvious one, but its consequences have generally been ignored by those working in natural language processing. A widely supported view of the natural language generation process, for example,[1] is that it involves *decision making under constraints*: constraints imposed by the context of language production, and by the linguistic resources available to the language producer. This view is rarely taken to its logical conclusion, however. There have been a few pieces of research that have looked at the computer production of language in time-limited situations, with the effects this has on what can be said and how it can be said (see, for example, [Hovy 1987], and some forthcoming work by Carletta and Caley at Edinburgh); but no current research takes account of the physical manifestation of the written word, and the constraints imposed by the *medium* of communication.

[1] The emphasis in this paper is on the process of natural language generation, but analogous concerns arise for natural language understanding too: our systems need to understand the impact of physical context as much as our generation systems need to be able to exploit it.

We focus here on one particular manifestation of these physical constraints: namely, that different means of signalling DISCOURSE STRUCTURE may be required in different circumstances, depending upon factors such as the general style of the document in question, and issues of space and layout. 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 such as *next*, *anyway*, *to get back to the point*;
- certain uses of punctuation markers, including commas, colons, semi-colons, dashes and parentheses; and
- graphical markers, including the use of paragraph breaks and itemized or enumerated lists.

Although there is little in the way of a rigorous treatment of the linguistic phenomena in question, the first of these is clearly uncontroversial. It is only very recently, however, that researchers have begun to take the others seriously. Nunberg [1990] argues that a pervasive, and often unconscious, 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. The view taken here is 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. The present 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.

3 Discourse Structure and Punctuation

3.1 Is Punctuation Rule-based?

We take the view here that many (but not all) uses of colons, semi-colons, parentheses, dashes and commas act as signals of discourse structure. Some uses of these punctuation marks are more amenable to a straightforwardly 'syntactic' characterisation; the use of commas to separate items in a list is one such case (as in the use of commas in *the flags were red, blue, and green*). Syntactic constraints also play a role when the primary function of a punctuation mark is at a higher level; thus, for example, there are quite clear rules about what happens when a digression bounded by dashes ends at the end of the sentence—as in this sentence. A quite sophisticated analysis of the syntactic constraints on the use of punctuation is presented in Nunberg [1990]; we will not address this issue here. Our aim is to look closer at the *semantic* aspects of the use of punctuation marks.

Any attempt to clarify the role of punctuation marks in indicating discourse structure is often met by an objection along the following lines: that there are no rules, and that the evidence for this is that we all use punctuation marks differently. It is certainly the case that our intuitions here are much less developed than in the case of straightforwardly syntactic phenomena. However, to the extent that an informal inquiry can establish any general tendency, my own view is that people's suspicions that they will disagree about the use of punctuation are generally mistaken: it turns out that the disagreements are more often about the *strength* of punctuation that is appropriate in a given circumstance (so that, for example, one person might find a comma sufficient where another would require a semi-colon), rather than disagreements based on *conflicting* use. Indeed, presented with examples of punctuation in context, there seems to be general agreement about which uses are clearly acceptable and which are not (consider replacing the comma in the current sentence with a colon and ask yourself whether you find the use of punctuation acceptable).

This leads me to suggest that, rather than there being no rules for the use of punctuation, there are indeed rules, but we rarely learn them in the way we learn rules of grammar. Since punctuation marks inhabit only the world of written language,[2] we have considerably less practice at using them than we do the rules of grammar proper, which we exercise in both written and spoken form. Perhaps because of their more limited use, any rules we might derive for the use of punctuation marks seem more like artificially stipulated conventions than do the rules of syntax; and it is partly in response to this perception, I suspect, that people feel uncomfortable with the notion that punctuation marks obey rules.

There is one place we can look for further guidance here. An examination of publishers' style guides suggests that there are a number of functions typically performed by punctuation marks. There is generally little disagreement between style guides—that's to say, they typically offer compatible but slightly different guidelines for the use of punctuation marks. Such guidelines are couched in terms like the following (from the *Chicago Manual of Style* [para 5.97]):

> Parentheses, like commas and dashes, may be used to set off amplifying, explanatory or digressive elements. If such elements retain a close logical relationship to the rest of the sentence, use commas; if the relationship is more remote, use dashes or parentheses ...

We are provided with the following example:

> Because the data had not been completely analysed—the reason for this will be discussed later—the publication of the report was delayed.

Descriptions of the use of punctuation marks in these terms are very common; a few more examples follow.

- *Hart's Rules for Compositors* [page 40] states the following:

  > ... whereas the semi-colon links equal or balanced clauses, the colon generally marks a step forward, from introduction to main theme, from cause to effect, premiss to conclusion, etc ...

[2] Nunberg [1990] argues convincingly against the view that punctuation marks are a written reflex of prosody.

- The *Chicago Manual of Style* [para 5.24] states that:

  > The comma indicates the smallest interruption in continuity of thought or sentence structure.

- The *Australian Government Manual of Style* [page 29] states that:

  > The semicolon indicates a pause or degree of separation greater than is marked by the comma but less than would justify a colon or full stop.

So, it's clear that these guides take the view that there are correct and incorrect ways of using punctuation. Each exhortation suggests that the punctuation marks should be used to make clear certain relationships in the text; relationships which are cast in terms similar to those often used in computational theories of discourse structure. However, characterisations like those we have just seen are obviously insufficiently well-specified to be of direct use in a computational system. To get closer to what's going on, we can try to come at the problem from a more theoretically motivated point of view.

3.2 Theories of Discourse Structure

A great deal has been written about the structure of discourse in the literature of computational linguistics and its sister disciplines, but the essential points remain quite invariant: theories of discourse structure generally concern themselves with trying to explain either the restrictions imposed in anaphoric reference imposed by the structure of discourse, or to explain the implicit relational propositions that inhabit a text and provide its coherence. For simplicity here, we will focus on two particular theories of discourse structure which are quite well known—Rhetorical Structure Theory (RST) and Grosz and Sidner's theory of discourse structure, which we'll refer to as GSDT—although many other theories would be equally appropriate choices.

Rhetorical Structure Theory is an essentially descriptive theory that aims to account for the rhetorical relations in texts. The theory makes the claim that, for most texts, a set of around twenty five relations suffices to label the relations that link segments (or SPANs) of text, regardless of their size. These relations are defined, relatively informally, in terms of preconditions on their use and in terms of their effects. This makes the theory potentially useful in the area of natural language generation (see, for example, [Hovy 1988; Moore and Paris 1989]), but too imprecise for use in natural language interpretation. The reader is directed to [Mann 1986] for a fuller description of the theory.

For our present purposes, it is sufficient to focus on a simple example. Consider the following two sentence discourse:

(1) a. I love to collect classic automobiles.
    b. My favourite car is my 1899 Duryea.

In RST terms, if we suppose that (1a) realises the proposition P_1 and (1b) realises the proposition P_2, then we would say that the relational proposition of ELABORATION holds between these two propositions:

(2) $\mathsf{elaboration}(P_1, P_2)$

In this example, P_1 is the NUCLEUS of the structure, and P_2 is the satellite; P_2 elaborates on P_1. This kind of analysis can be applied to text spans of any size, with the result for any given text being a tree of relations that characterises the rhetorical structure of the text as a whole.

Grosz and Sidner's Theory of Discourse Structures (henceforth GSDT), on the other hand, has different concerns: in particular, it attempts to explain how the intentional structure (the goals and purposes) underlying a text constrains the process of anaphora resolution.[3] The basic idea here is that a text has a discourse structure which mirrors the intentional structure that underlies the text; corresponding to the text as a whole there will be some discourse purpose, and the text can be decomposed hierarchicaly into segments, each of which corresponds to a discourse segment purpose which contributes to the purpose of the larger discourse segment of which it is a part. The hierarchy of discourse segments is then claimed to give rise to particular constraints on the scope of anaphoric expressions. Relations betwen discourse segment purposes are of two types: either one purpose DOMINATES another, in which case the latter purpose contributes to the former, or one purpose SATISFACTION-PRECEDES the other, so that it serves as a precondition for the second purpose. Grosz and Sidner are keen not to adopt a collection of semantic or rhetorical relations in the way that RST does, arguing that any such set must be open-ended, and viewing this as an undesirable characteristic.

According to GSDT, in the text in Figure 1, the hearer will have no difficulty in resolving the referent of the noun phrase *the tent* in utterance 14 as being the tent mentioned in utterance 2: the more recently mentioned tent mentioned in utterance 7 will no longer be available for anaphoric reference because the discourse segment (labeled here DS1) in which it is mentioned has been completed, and thus removed from the focus of attention. The reader is referred to Grosz and Sidner [1987] for a fuller exposition of the theory; see Dale [1988] for a discussion of some concerns about the central claims.

3.3 Some Data

Armed with the key notions of theories such as these, it is instructive to look at some examples of the use of punctuation. Consider the natural interpretations of each of the following examples (borrowed from Nunberg [1990]):

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

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

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

We might suppose each sentence here to embody the same two propositions, glossed as follows:

P_1: Some person reported the decision.

P_2: We were forbidden to speak with the chairman directly.

In each case, however, it seems plausible to suggest that the different punctuation marks are saying something different about the relationships between the propositions. We can

[3] The theory attempts to explain much more than simply anaphora resolution, but this is the most relevant aspect here, and arguably the most generally useful.

DS0	1.	A:	I'm going camping next weekend. Do you have a two-person tent I could borrow?
	2.	B:	Sure. I have [a two-person backpacking tent].
DS1	3.	A:	The last trip I was on there was a huge storm.
	4.		It poured for two hours.
	5.		I had a tent, but I got soaked anyway.
	6.	B:	What kind of a tent was it?
	7.	A:	A tube tent.
	8.	B:	Tube tents don't stand up well in a real storm.
	9.	A:	True.
	10.	B:	Where are you going on this trip?
	11.	A:	Up in the Minarets.
	12.	B:	Do you need any other equipment?
	13.	A:	No.
	14.	B:	Okay. I'll bring [the tent] in tomorrow.

Figure 1: The Tent Example

characterise these differences in terms of a theory like Rhetorical Structure Theory. Examples (3) and (4) seem to exhibit instances of ELABORATION and CAUSE respectively, and example (5) could express either of these rhetorical relations: in (3), the fact that the speaker and others are forbidden to speak with the chairman directly seems to be the content of the reported decision, whereas in (4), it seems that the speaker and others being forbidden to speak with the chairman directly is the reason that some individual (referred to here as *he*) reported the decision.

This difference in the inferred relations between the propositions in the text suggests that punctuation markers at least play some role in the interpretation of particular rhetorical relations. The mapping is not straightforwardly one-to-one, however, as (5) shows: at the very least, different relations may be realized by means of the same punctuation markers.

3.4 The Nature of the Underdetermination

It should not be particularly surprising 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 rhetorical relations.

We are then led to enquire after the precise nature of this underspecification. 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.

However, 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]).

(6) 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 expressed here between the two constituent propositions as being one of ELABORATION or one of ANTITHESIS: on the one hand, we may be being told that John's view of the book is an example of the general sentiment described in the first clause, while on the other hand, the writer may be contrasting John's view of the book with that of some other people. A given reader's interpretation of the punctuation mark, if it is playing a role here at all, will no doubt be influenced by the reader's attitude towards post-modern criticism.

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-precedence (SP) relations, but say nothing about the particular *semantics* of the relations.

3.5 Discourse Structure and Grain Size

Another question which arises when we consider punctuation from the perspective of theories of discourse structure is that of determining just what the units are that are being related. In RST, rhetorical relations are seen to hold between text spans, which can be of any size, but are ultimately grounded out as single clauses. In GSDT, discourse segments are less well-defined, although the constraints the theory imposes on anaphora require that discourse segments typically consist of two or more clauses (if all single clauses were discourse segments, pronominal reference to entities mentioned in the previous clause would not be possible). Some punctuation uses suggest, however, that we might also want to look *within* individual clauses when attributing discourse structural relations to a text.

Notice that the writer is often free to use structure indicators from any of the three systems we identified at the beginning of the paper. 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

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

or by using more graphical devices, as in:

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

Given the simple domain, this last example is somewhat artificial, but the general idea should be clear; graphical structuring 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 (9) over (10):

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

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

Again, the role of the punctuation marks is crucial here. However, as we noted above, 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 (7) and (8), and the need to view the non-restrictive relative clause in (10) 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.

4 Towards a Taxonomy of Punctuation Function

There are clearly many unanswered questions here. As a first step towards determining just what the role of punctuation is in indicating discourse structure, we present in this section a simple taxonomisation of the specific functions performed by punctuation marks, derived from a close examination of the kinds of constraints indicated by style guides. We suggest that punctuation marks can indicate:

- degree of rhetorical balance: i.e., the relative importance of juxtaposed elements;
- aggregation: i.e., the relative closeness and distance of juxtaposed material; and
- particular rhetorical relations: some punctuation marks seem to play a role in indicating what semantic or rhetorical relations hold between juxtaposed elements.

4.1 Rhetorical Balance

One common characterisation of the differences between punctuation marks is that of the relative weight of the elements connected by the marks. Recall from earlier *Hart's Rules*' claim that:

> ... whereas the semi-colon links equal or balanced clauses, the colon generally marks a step forward, from introduction to main theme, from cause to effect, premiss to conclusion, etc ...

This distinction is very reminiscent of the distinction within RST between NUCLEI and SATELLITES: most relations in RST relate some nuclear material to some satellite material, with the

satellite material being less important (to the extent that it can be deleted without the sense of the text being affected, whereas if a nucleus is deleted, an incoherent text results). A few relations, however, are multi-nucleur, in the sense that neither of the related text spans is more important than the other. Further analysis is required to see whether this difference corresponds to the kind of rhetorical balancing suggested by the style rule above.

4.2 Aggregation Markers

Another function of the different punctuation marks is to indicate how closely together material is related, and to indicate differing degress of relatedness. Thus, as we saw earlier, the *Chicago Manual of Style* states that:

> The comma indicates the smallest interruption in continuity of thought or sentence structure.

and the *Australian Government Manual of Style* states that:

> The semicolon indicates a pause or degree of separation greater than is marked by the comma but less than would justify a colon or full stop.

It is relatively straightforward to see how this notion of aggregation would fit neatly into the hierarchy of discourse segment purposes that GSDT sees as underlying a text, although again the detailed ramifications of this need further working out.

Related to this, most sources agree that, used in pairs, different punctuation marks indicate different degrees of embedding: For example, the *Chicago Manual of Style* [para 5.97] states that:

> Parentheses, like commas and dashes, may be used to set off amplifying, explanatory or digressive elements. If such elements retain a close logical relationship to the rest of the sentence, use commas; if the relationship is more remote, use dashes or parentheses ...

For example:

(11) Because the data had not been completely analysed—the reason for this will be discussed later—the publication of the report was delayed.

4.3 Rhetorical Markers

Other descriptions of the function of punctuation marks point towards particular rhetorical relations.

Colons: Style guides seem to unanimously agree that colons are used:

- to emphasise a sequence in thought;
- to illustrate, amplify or explain.

For example:

(12) Many of the policemen held additional jobs: thirteen of them, for example, doubled as cab drivers.

Commas: According to some style guides, commas are used:

- to emphasise a point or subtle distinction;
- to indicate contrast (antithesis).

For example:

(13) As there were wicket-keepers before Blackham, so there were labour unions before the gold discoverers.

(14) The fool wonders, the wise man asks.

This categorisation is, of course, very impressionistic. Further work needs to be done to determine whether the uses of punctuation can be so straightforwardly tied to the theoretical notions underlying current theories of discourse structure as these observations suggest.

5 Conclusions

This paper has raised a number of questions for further research:

- Are punctuation marks just like lexical markers of discourse structure?
- How does the space of punctuation markers map onto the taxonomy of relations? What does the use of punctuation markers tell us about the structure of the space of relations?
- How are the choices between graphical, punctuation and lexical markers made?
- What is the atomic unit of discourse structure? Where does discourse structure stop? Is there really a discourse/syntax boundary?

It seems clear that punctuation plays an important role in helping the reader infer the structure of a discourse, and so should be used and understood appropriately by natural language processing systems. Answering the above questions may give insights that will further develop the computational study of discourse structure.

References

R Dale [1988] The generation of subsequent referring expressions in structured discourses. Pages 58–75 in M Zock and G Sabah (eds), *Advances in Natural Language Generation: An Interdisciplinary Perspective*, Volume 2. Pinter Publishers Ltd, London.

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 [1987] Generating Natural Language Under Pragmatic Constraints. PhD Thesis, Department of Computer Science, Yale University.

E Hovy [1988] Planning coherent multisentential text. Pages 163–169 in *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, State University of New York at Buffalo, Buffalo, N.Y., 7-10 June, 1988.

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.*

W C Mann and S Thompson [1986] Rhetorical Structure Theory: Descripton and Construction of Text. Report RS-86-174, Information Sciences Institute, University of Southern California.

J D Moore and C L Paris [1989] Planning Text for Advisory Dialogues. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, University of British Columbia, Vancouver, B.C., 26–29 June, 1989.

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.