Information-Based Aspects of Punctuation

Bilge Say and **Varol Akman**
Department of Computer Engineering and Information Science,
Bilkent University,
Bilkent, 06533 Ankara, Turkey
{say,akman}@bilkent.edu.tr

Abstract

We offer a preliminary account of the information-based aspects of punctuation marks. We give our initial treatment within the Discourse Representation Theory and its segmented version. We hypothesize that this work will be useful in classifying the informational contributions of punctuation marks and bringing them to bear on the semantic characterization of written discourse.

1 Introduction

Recent linguistic works have attempted to produce systematic characterizations of punctuation marks descriptively.[1] Nunberg (1990) shows how punctuation is a linguistic system on its own and devises a "text-grammar" for this purpose using mechanisms of conventional, or "lexical" grammars. Based on his work, several researchers integrated punctuation marks into the NLP systems (Briscoe, 1994; Jones, 1994; White, 1995). We want to add on top of the previous work a formal characterization of the information that punctuation marks bring to the discourse, semantically and pragmatically, within or above (grammatical) sentence level.

2 Punctuation and Information

We take information as the propositional content of a sentence which constitutes a contribution to reader's knowledge store (as used in *information packaging* by Vallduví (1992)). We show how punctuation marks can provide informational cues via various channels in Figure 1.

Punctuation marks play various informational roles in natural language discourse. They can have a morphological role such as in *anti-feminist*, a delimiting role such as in *Jones, my brother, came yesterday*, or a separating role such as in *two bottles of wine, three cans of beer*. They can also have distinguishing roles such as usage of capital letters for proper names. These roles sometimes serve to resolve ambiguities, e.g., *new, regular time for Tai-Chi classes* as compared to *new regular time for Tai-Chi classes*. If our intended meaning is to announce classes with a fixed schedule, the second construct would be ambiguous. As in this example, some of these roles of punctuation may have semantic functions. Our claim is that they can even change the analysis of discourse. In fact, various punctuation marks operate above sentence level connecting independent clauses that can function as stand-alone sentences. In addition, these connections result in special effects such as elaboration. In this respect, discourse usage of punctuation marks are similar to relations in Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) relations as noted by Dale (1991). RST involves characterizing coherence relations that hold between arbitrarily long units of text.

There is also an interaction between punctuation and intonation in bringing out the informational cues (Bolinger, 1989; Chafe, 1988). As we are going to deal with written language, we will not delve into that further.

We will initially concentrate on *structural* marks in English as Meyer (1983) suggests, studying only those marks that act on units not larger than the orthographic (written) sentence (thus no paragraphs) and not smaller than the word (thus no hyphens or apostrophes).

3 Punctuation in Discourse

Our aim is to be able to capture the effects of punctuation within a formal framework. A suitable choice looks like the Discourse Representation Theory (DRT) by Kamp and Reyle (1993) which integrates current

[1] A survey is available in (Say, 1995).

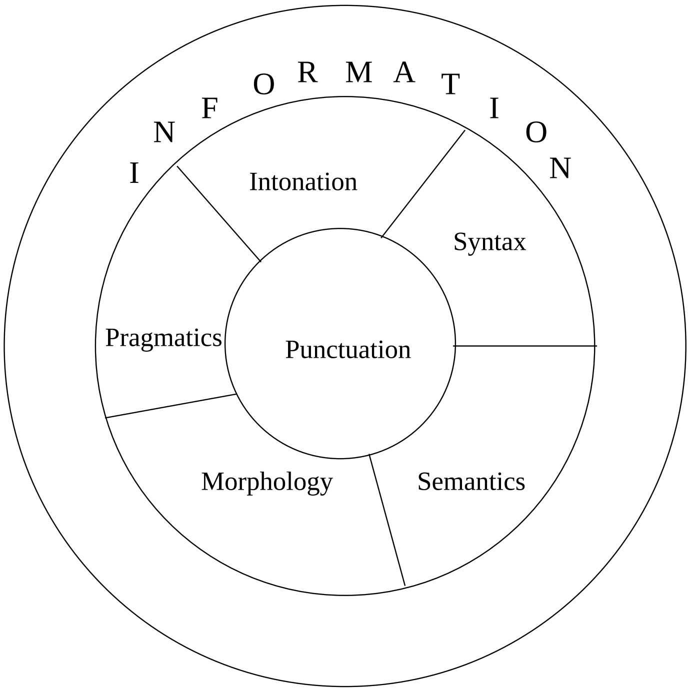

Figure 1: Punctuation as Information

approaches in a semantic theory. The aim of DRT has been stated as "providing a systematic specification of the truth conditions of multi-sentential discourses and texts" (Kamp and Reyle, 1993). To be able to do this, representational devices called Discourse Representation Structures (DRSs) are built up while the discourse is being interpreted. DRT has not only well-developed accounts for anaphora, quantification, tense, etc., but also applicability in a strong computational sense, which will be necessary for our work. However it lacks, in its bare bones version (Kamp and Reyle, 1993), constructs that deal with the structure and the relations of the discourse, which are required for certain usages of punctuation. Such constructs are provided by Asher (1993) within another theory he presents for discourse structure for analyzing abstract entity anaphora. The structure and the segmentation of discourse may help to choose antecedents for anaphoric reference. The basic entities at this level are called *segmented DRSs* (SDRSs) by Asher. They are imposed on the logical structure created by DRSs by relating DRSs with discourse relations, which act as conditions for SDRSs. Built incrementally as DRSs, a unit of information is defined to be a *constituent*. Asher takes a basic constituent to correspond to a sentence ended by a full-stop as default, though this can be overridden by clauses or longer stretches of text where required. We will investigate cases triggered by punctuation marks to force such a processing in the level of subsentence phenomena. Asher uses a subset of relations from RST (Mann and Thompson, 1987) and other discourse structure theories for his purposes. He designates certain relations as affecting the hierarchical structure of the text. In dealing with parenthetical constructs such as those implied by dashes we will have to make use of this hierarchical structure. Also important are *parallelism* and *contrast* that involve pairing structurally similar objects according to whether they are semantically similar or dissimilar, respectively (viz. in usages of semicolons, etc.).

Considered below are several types of punctuated sentences that influence the semantics and the pragmatics of the discourse. We briefly comment on them to show how they can be dealt with DRSs or SDRSs. (To avoid cluttering, tense and various other information have in general been omitted from the following DRSs.)

(1) a. Tom has two cats that once belonged to Fred, and Sam has one.
b. Tom has two cats, which once belonged to Fred, and Sam has one. (McCawley, 1981, p. 103)

(1a) implies that Sam has a cat that once belonged to Fred whereas (1b) implies that Sam has a cat but there is no information as to whether it once belonged to Fred. This kind of construct can straightforwardly be dealt with plain DRSs as shown in Figure 2.

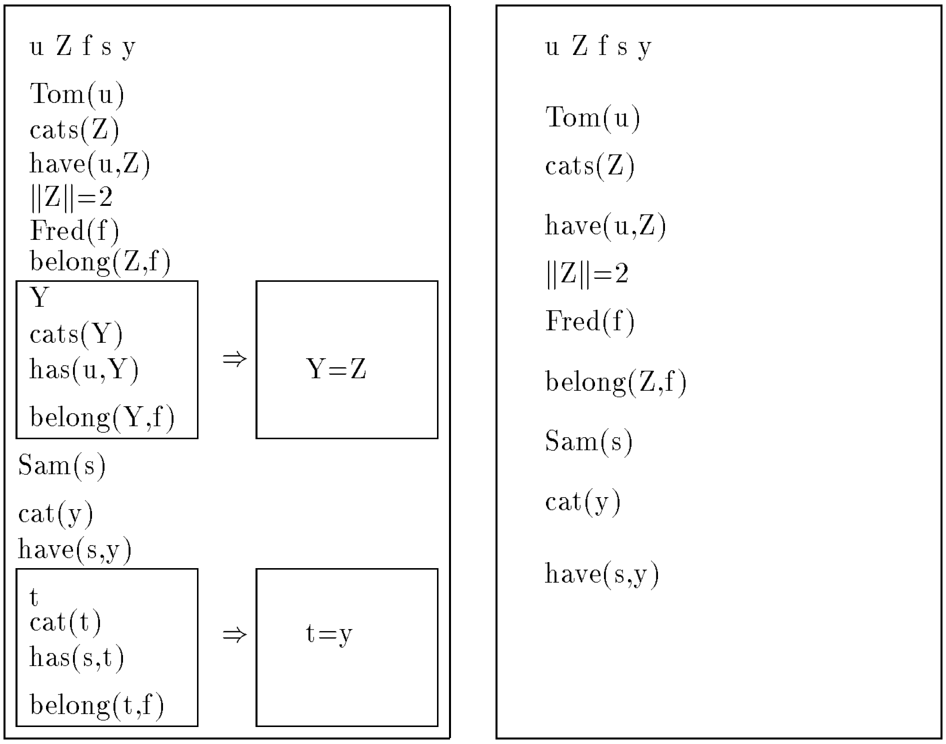

Figure 2: DRSs for (1a) and (1b)

(2) a. Jane, and Joe and Sue write books on England. If her books are best-sellers then they are jealous.

b. Jane and Joe, and Sue write books on England. If her books are best-sellers then they are jealous.

DRS for (2a)	DRS for (2b)
x y z X Z W	x y z X Z W
Jane(x)	Jane(x)
Joe(y)	Joe(y)
Sue(z)	$X = x \oplus y$
$X = y \oplus z$	Sue(z)
$Z = x \oplus X$	$Z = X \oplus z$
books on England(W)	books on England(W)
write(Z,W)	write(Z,W)
[u; x's books(u); $u \in W$; be(u,best-sellers)] $\Rightarrow$ [X are jealous]	[u; z's books(u); $u \in W$; be(u,best-sellers)] $\Rightarrow$ [X are jealous]

Figure 3: DRSs for (2a) and (2b)

The exact position of the comma in the first sentence changes the resolution of pronominal anaphora in the second sentence, which is identical in both pieces of discourse. (2a) will have *her* attached to Jane and *they* to Joe and Sue, whereas (2b) will have *her* attached to Sue and *they* to Jane and Joe. This can also be dealt with plain DRSs as shown in Figure 3.

(3) John — his brother also an athlete — won the university medal easily. He is an ambitious guy.

In (3) *he* must be resolved to John, not to his brother, as the material within dashes is parenthetical. To deal with such sentences we have to modify the SDRS construction and take advantage of discourse structure. Here, *Parenthetical* is a new relation in that respect. The *Elaboration* relation implies that the first constituent of the relation is an elaboration for the second. The relevant SDRS is in Figure 4.

(4) She looks right, he looks left; she smiles, he frowns; she clasps her hand around her knee, he clasps his around his head. (Bolinger, 1989, p. 183)

In (4) there is not only a temporal sequence but also a relation of causality between the subsentences separated by commas. We can make use of the *Parallel* relation here, as shown in Figure 5.

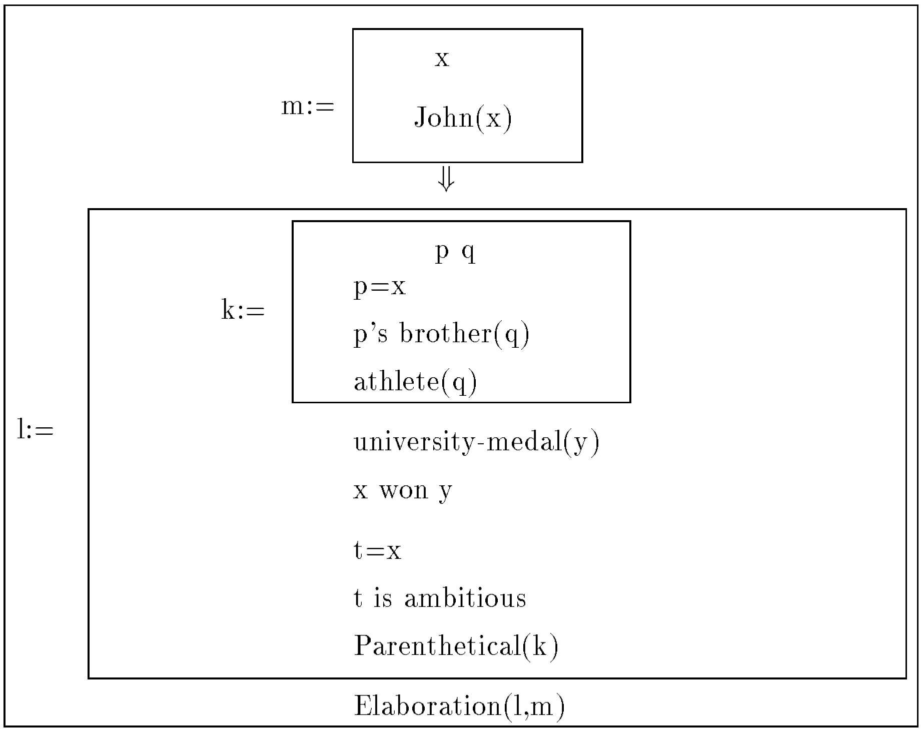

Figure 4: SDRS for (3)

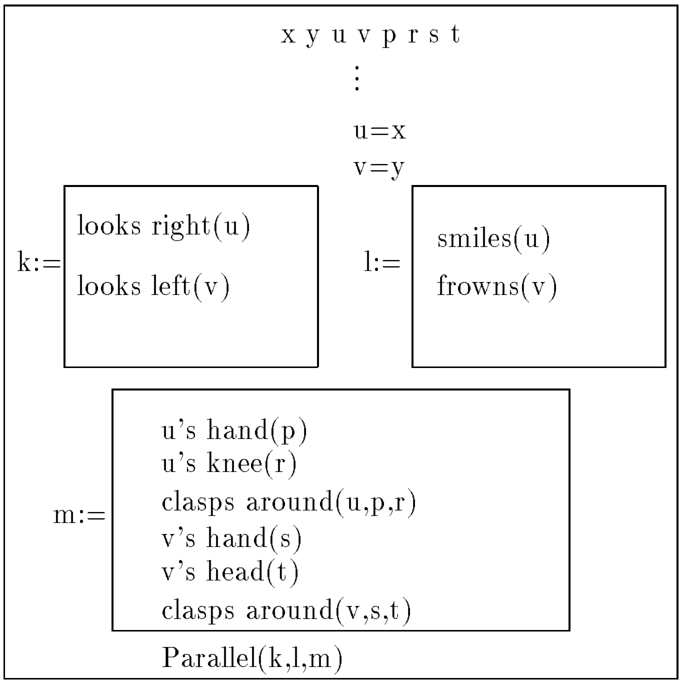

Figure 5: SDRS for (4)

(5) Today, John went to school. He has been hospitalized for a year. (Dawkins, 1995, p. 537)

x e y u v s John(x) today(e) school(y) k: e-go(x,y) FOCUS(k)=e u=x hospital(v) s-in(u,v) dur(s)=1 after(e,s)

Figure 6: DRS for (5)

In (5) the comma coincides with an intonation group boundary to indicate focus. We understand that John has been unable to go to school for a year so *today* is a special day. Neither DRSs or SDRSs can show such information structure so we have to introduce a new construct. The FOCUS function shows the focus of the relevant sentence as in Figure 6.

(6)
a. He reported the decision: we were forbidden to speak with the chairman directly.
b. He reported the decision; we were forbidden to speak with the chairman directly. (Nunberg, 1990, p. 13)

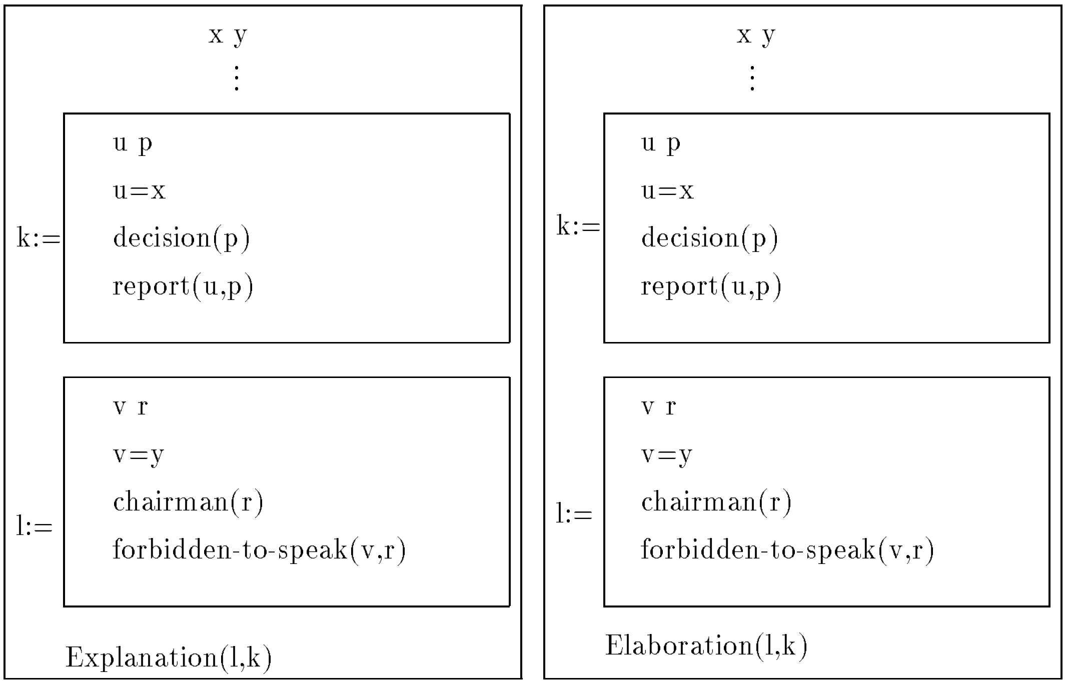

Figure 7: SDRSs for (6a) and (6b)

(6a) takes the decision to be the ban of spoken interaction with the chairman. (6b), on the other hand, is more inclined to indicate that because of the ban, another person, not the chairman reported the decision. This distinction can be captured by changing the SDRS building algorithms and directing the punctuation mark to the appropriate relation (the constituent l being an *explaining* k in (6a) and *elaborating* it in (6b)) as shown in Figure 7. However, (6b) can also be ambiguous between the preferable reading and the meaning of (6a). Resolving such an ambiguity without contextual information is a problem.

(7) a. The great days faded. The end is in sight.
b. The great days faded; the end is in sight (Dawkins, 1995, p. 541)

Apparently, (7b) has more emphasis and linkage than (7a) but this can also be a matter of style. Whether it is worthwhile to capture the relation between the parts of (7b) using SDRSs is an open question.

As can be seen, the underdetermination of punctuation marks present problems as, sometimes, two marks can be used interchangeably without a marked distinction in the meaning. Other times, the distinction can only be determined within an appropriate context or can depend on personal style. We must concentrate on consistent usage as much as possible by choosing certain genres of text so as to limit the effects of the problem. Overdetermination is also present, for the marks can provide simultaneous cues at the same time. Dealing with overdetermination is less problematic as different layers within the theory can be made to accommodate different cues.

Our initial assessment is that the model theory of DRSs does not have to be affected a lot since the relations envisaged, whether they are between constituents or between a subDRS and the DRS itself, are actually additional information to the existing DRSs and have to be processed that way with the additional operators introduced and defined. We have yet to find ways to fine-tune and integrate these ideas in the standard theory (Kamp and Reyle, 1993).

4 Conclusion

Most of the uses of punctuation marks can be rightly seen in an information-based context. Obtaining an adequate formalism to capture information cues may also positively interact with studies in related phenomena such as discourse markers (Schiffrin, 1987) or intonation. We have given examples of how such information-based punctuation marks can be treated within Discourse Representation Theory. We aim to extend our coverage to a fuller set of uses of various punctuation marks.[2] After such a treatment, we hope

[2] Related work that will provide useful data is being conducted by two students in the form of corpus analysis (Sampson, 1995) of punctuation mark usage.

to make it a worthwhile endeavour to make the results apply in a computational setting. This might involve extending a suitable DRT implementation to integrate the effects of punctuation mark usage.

5 Acknowledgment

This work was partially supported by a NATO Science for Stability project grant TU-LANGUAGE.

References

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse.* Studies in Linguistics and Philosophy. Kluwer Academic Publishers, Dordrecht, Netherlands.

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Stanford University Press, Stanford, California.

Briscoe, Ted. 1994. Parsing (with) Punctuation. Technical report, Rank Xerox Research Centre, Grenoble, France.

Chafe, Wallace. 1988. Punctuation and the Prosody of Written Language. *Written Communication*, 5(4):395–426.

Dale, Robert. 1991. Exploring the Role of Punctuation in the Signalling of Discourse Structure. In *Proceedings of a Workshop on Text Representation and Domain Modelling: Ideas from Linguistics and AI*, pp. 110–120. Technical University Berlin.

Dawkins, John. 1995. Teaching Punctuation as a Rhetorical Tool. *College Composition and Communication*, 46(4):533–548.

Jones, Bernard. 1994. Exploring the Role of Punctuation in Parsing Natural Language. In *Proceedings of COLING '94*, pp. 421–425, Kyoto, Japan.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic.* Parts 1 and 2. Kluwer Academic Publishers, Dordrecht, Netherlands.

Mann, William C. and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report RS-87-190, USC Information Sciences Institute, Marina Del Rey, California.

McCawley, James D. 1981. The Syntax and Semantics of English Relative Clauses. *Lingua*, 53:99–149.

Meyer, Charles F. 1983. *A Linguistic Study of American Punctuation.* Ph.D. thesis, University of Wisconsin-Milwaukee.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation.* CSLI Lecture Notes No. 18. CSLI Publications, Stanford, California.

Sampson, Geoffrey. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Oxford University Press, Oxford.

Say, Bilge. 1995. An Information-Based Approach to Punctuation. Ph.D. Proposal, Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey. Available on the WWW as follows: `http://www.cs.bilkent.edu.tr/~say/bilge.html`.

Schiffrin, Deborah. 1987. *Discourse Markers.* Cambridge University Press, Cambridge, UK.

Vallduví, Enric. 1992. *The Informational Component.* Outstanding Dissertations in Linguistics. Garland Publishing, New York.

White, Micheal. 1995. Presenting Punctuation. In *Proceedings of the Fifth European Workshop on Natural Language Generation*, pp. 107–125, Leiden, Netherlands.