An Information-Based Treatment of Punctuation in Discourse Representation Theory

Bilge Say and Varol Akman

Bilkent University

Abstract

Punctuation has so far attracted attention within the linguistics community mostly from a syntactic perspective. In this paper, we give a preliminary account of the information-based aspects of punctuation, drawing our points from assorted, naturally occurring sentences. We present our formal models of these sentences and the semantic contributions of punctuation marks. Our formalism is a simplified analogue of an extension—due to Nicholas Asher—of Discourse Representation Theory.

Introduction

Punctuation marks have not been studied much by linguists apart from a prescriptive angle, until recently. Similarly, most Natural Language Processing (NLP) systems do not take punctuation marks into account (except for the period and the spacing). Some linguistic works have attempted to produce systematic characterizations of punctuation marks descriptively.

Levinson (Levinson 1985) emphasizes the distinction between the orthographic sentence and the grammatical one, as well as explaining how punctuation marks bind entities according to their informational links. Meyer (Meyer 1987) gives a classification of punctuation marks (in American usage) according to their functions and studies the realization of those functions. Nunberg (Nunberg 1990) shows how punctuation is a linguistic system on its own and devises a "text-grammar" for this purpose, using mechanisms of lexical grammars.

Based on Nunberg's seminal work, several researchers have tried to integrate punctuation marks into their NLP systems, frequently using a syntactic point of departure (Briscoe 1996, Jones 1997, White 1995). Osborne (Osborne 1996) has shown the improvement—in a combined model-based and data-driven grammar learning approach—obtained as a result of using knowledge of punctuation to enhance the grammar, although his assumption of punctuation marks of all kinds attaching to maximal projection phrases is too strong. Kettunen (Kettunen 1996) has underlined the need for syntactic and semantic contexts in high-level "typographical spell checking" in his somewhat prescriptive study.

The overall, long-term goal of our research in punctuation is to contribute toward the construction of a systematic and principled theory of punctuation and human information processing vis-à-vis punctuation marks. More specifically, we want to add on top of the existing useful works a formal characterization—formulated in a contemporary semantic theory (Kamp and Reyle 1993)—of the information that punctuation contributes to the discourse, semantically and pragmatically, within or above the grammatical sentence level.[1]

Punctuation and Discourse Representation Theory

Punctuation marks play various roles in natural language texts. 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* versus *new regular time for Tai-Chi classes*. If our intended meaning is to

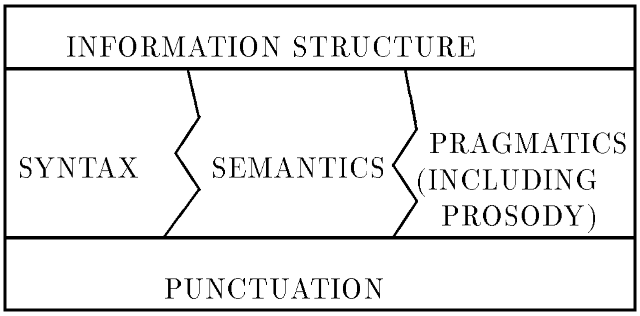

Figure 1: *Effects of punctuation to information structure*

announce classes with a fixed schedule, the second description would be ambiguous. As in the last example, some of the roles of punctuation may have semantic roots. Our claim is that punctuation may even change the analysis of discourse. In fact, various punctuation marks operate above the sentence level, connecting independent clauses that can function as stand-alone sentences. In addition, these connections can deliver special effects such as "elaboration" (Mann and Thompson 1987).

Vallduví's (Vallduví 1992) treatment of *information packaging* may be used to explain, in a rather abstract sense, various effects of punctuation marks (cf. Figure 1). By information packaging, he means the non-truth-conditional meaning of a sentence and how the latter is brought about. Information is defined as the propositional content which constitutes a contribution of knowledge to a reader's knowledge store. Vallduví gives the following simple example (Vallduví 1992: 2):

(1a) He hates broccoli.

(1b) Broccoli he hates.

Sentences (1a) and (1b) are clearly truth-conditionally equivalent but they say what they claim about *he* (a particular male) in different ways. Vallduví devises a scheme that could account for such differences in information packaging. He cautions that the way information packaging is realized linguistically (i.e., by means of intonation, syntax, or morphology) may differ from language to language. Following a common trend in the punctuation research community, we concentrate on the structural punctuation marks and take the combined effects of syntactic, semantic, and (to a certain extent) pragmatic uses in order to explain the value punctuation marks add to the information structure of a sentence. By *structural*, we mean those marks that act on units not larger than the orthographic sentence (thus no paragraphs) and not smaller than the word (thus no hyphens or apostrophes).[2]

While Vallduví's framework is very instructive, a more suitable and precisely defined medium to realize our goal is the *Discourse Representation Theory* (DRT) (Kamp and Reyle 1993). This is a formal proposal to integrate the current approaches to discourse in a full-fledged semantic theory. The aim of DRT has been stated as providing a systematic specification of the truth conditions of a given multi-sentential discourse (or text). To be able to do this, representational devices called the *Discourse Representation Structures* (DRSs) are incrementally built while the discourse is being interpreted in a top-down fashion.

As a somewhat superficial example to the "discourse effects" of punctuation, consider:

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

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

In both fragments, the exact position of the comma alone controls the proper resolution of pronominal anaphora. Suitable "triggering configurations" (Kamp and Reyle 1993) will lead to different structures within DRT: in (2a) we have *her* attached to Jane and *they* to Joe and Sue, whereas in (2b) we have *her* attached to Sue and *they* to Jane and Joe. This difference can be handled with plain DRSs—enriched with temporal and aspectual information, when necessary—as shown in Figure 2.

As for the effects of restrictive and nonrestrictive clauses, example (3a) below implies that Sam has a cat that once belonged to Fred whereas (3b) implies that Sam has a cat but there is no information as to whether it once belonged to Fred (both sentences taken from (McCawley 1981: 103)). This semantic distinction can straightforwardly be dealt with plain DRSs (cf. Figure 3).[3]

(3a) Tom has two cats that once belonged to Fred, and Sam has one.

(3b) Tom has two cats, which once belonged to Fred, and Sam has one.

$$
\begin{array}{|l|}
\hline
\qquad\qquad x\ y\ z\ X\ Z\ W \\
\hline
\text{Jane}(x) \\
\text{Joe}(y) \\
\text{Sue}(z) \\
X = y \oplus z \\
Z = x \oplus X \\
\text{books-on-England}(W) \\
\text{write}(Z,W) \\
\begin{array}{|l|}
\hline
\qquad U \\
\hline
x\text{'s books}(U) \\
U \subseteq W \\
\text{be}(U,\text{best-sellers}) \\
\hline
\end{array}
\Rightarrow \boxed{\text{be-jealous}(X)} \\
\hline
\end{array}
\qquad
\begin{array}{|l|}
\hline
\qquad\qquad x\ y\ z\ X\ Z\ W \\
\hline
\text{Jane}(x) \\
\text{Joe}(y) \\
\text{Sue}(z) \\
X = x \oplus y \\
Z = X \oplus z \\
\text{books-on-England}(W) \\
\text{write}(Z,W) \\
\begin{array}{|l|}
\hline
\qquad U \\
\hline
z\text{'s books}(U) \\
U \subseteq W \\
\text{be}(U,\text{best-sellers}) \\
\hline
\end{array}
\Rightarrow \boxed{\text{be-jealous}(X)} \\
\hline
\end{array}
$$

Figure 2: *DRSs for (2a) and (2b), respectively*

Asher's Extension and Its Adaptation to Modeling Punctuation

The simple examples in the preceding section give a taste of DRT capturing some semantic effects of punctuation. Still, more involved requirements for modeling punctuation effects are not readily expressible in standard DRT, and we have to define additional constructs. Several such constructs are provided by Asher (Asher 1993) within another theory he develops for analyzing abstract entity anaphora in discourse. According to Asher, the structure and the segmentation of discourse may help to determine the antecedents of abstract anaphoric references. To see an example of what is meant by "abstract," consider the following paragraph (excerpted from (Asher 1993: 346)):

(4) The Ashers were predictably short of groceries the day of the party. Nicholas Asher went out to get some, got lost and arrived back only after the party had ended. Because of this, the committee made sure that the Ashers never gave a party for the Society again.

Here, *this* anaphorically picks up a certain sum of events. Just which events are available for anaphoric reference is a result of the discourse structure. The basic entities at this stage are called the *Segmented* DRSs (SDRSs). They are imposed on the logical structure created by the DRSs by associating the DRSs with discourse relations, which act as conditions for SDRSs.

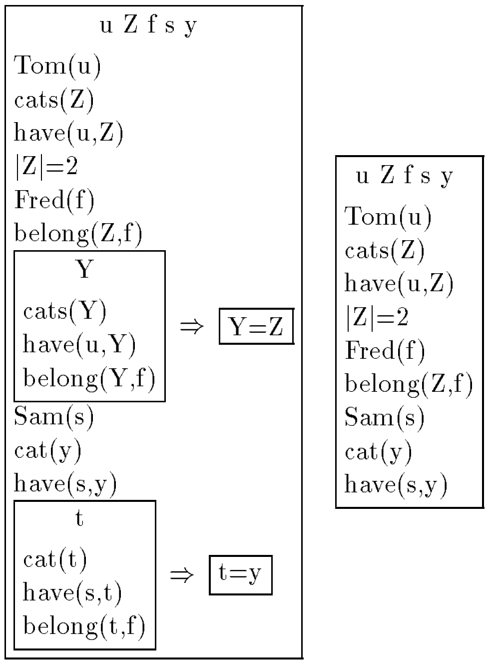

Figure 3: *DRSs for (3a) and (3b), respectively*

Asher takes, by default, each basic constituent of an SDRS to correspond to a sentence ended by a full-stop; this can clearly be overridden by clauses or longer stretches of text where required. When the SDRS structure is updated, the constituents are revised for possible anaphora resolution, and the truth condition changes. An important rule is that the coherence of the discourse must be maintained as new constituents are being attached.

Asher uses a subset of relations from the *Rhetorical Structure Theory* (RST) (Mann and Thompson 1987) and some other discourse structure theories for his purposes. He divides the relations he uses in his theory of SDRSs into two (viz. rhetorical and coherence relations) according to how they relate sentences and contribute to the truth conditions. More importantly (at least for our aims), he also classifies the relations according to whether they affect the hierarchical structure of the text. *Summary* and *topic* are examples of such structural relations that dominate other constituents. Also important are the relations *parallel* and *contrast*, which involve pairing structurally similar objects according to whether they are semantically similar or dissimilar, respectively.

Asher's theory of SDRSs has been very influential in our work. In a way, we had to incorporate intra-sentential punctuation phenomena to his theory. (His theory obviously deals with inter-sentential discourse phenomena.) While the exact definitions of the relations used in our punctuation-oriented study may be found in (Say 1995), the following illustrative examples may offer a glimpse of this study.

In example (5) (taken from (Meyer 1986: 81)), the sentence-final dash indicates a *result* of the previously accumulated eventualities and does not directly represent a parenthetical occurrence. The state expressed after the dash follows from the subsentences before it. The resulting SDRS is given in Figure 4.[4]

(5) She had cried, she had implored, she had been miserable at this refusal, and finally he had relented—and now how happy she was, how expectant!

In contrast to the previous sentence, example (6) below (adapted from (Ehrlich 1992: 80)) also has a sentence-final dash which does indicate a parenthetical. As a solution to this under-determination problem, we may need to resort to heuristics, founded on corpus analysis of punctuation mark usage, because rules alone do not suffice to determine the effect of the dash. The relevant SDRS for (6) is also depicted in Figure 4.

(6) Now, I tell you the entire story—but first you have another cup of coffee.

In examples (7) (taken from (Meyer 1986: 83)) and (8) (taken from (Ehrlich 1992: 81)), the parenthetical remarks clarify the parts of a plural discourse referent (*a variety of environmental supports* and *the problem*, respectively). These remarks are important for the discourse when such a clarification is sought by the reader. The relevant SDRSs are shown in Figures 5 and 6, respectively.

(7) Simultaneously, a variety of environmental supports—a calm but not too motherly homemaker, referral for temporary economic aid, intelligent use of nursing care, accompaniment to the well-baby clinic for medical advice on the twin's feeding problem—combined to prevent further development of predictable pathological mechanisms.

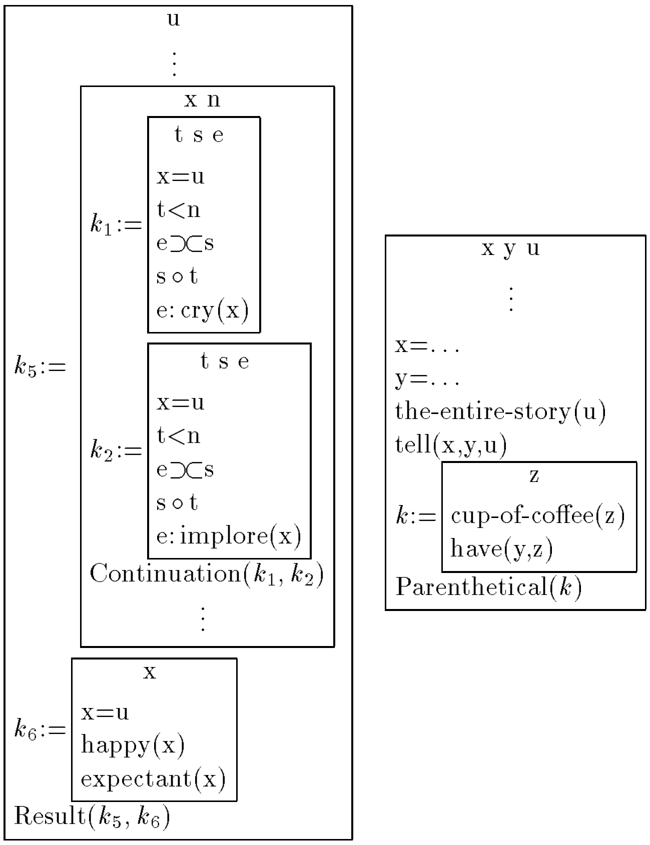

Figure 4: *SDRSs for (5) and (6) [some details omitted]*

(8) The problems—unemployment and inflation—perplex economists and mystify the public.

In example (9) (taken from (Dawkins 1995: 537)), 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; therefore *today* is a special day. Neither a plain DRS nor an SDRS can show such an information structure: we simply have to introduce a new construct. The *focus* relation can be used to this end, as Figure 7 illustrates.

(9) Today, John went to school. He had been hospitalized for a year.

We can now consider semantically contributing instances of semicolons and colons. In example (10) (taken from (Quirk et al. 1972: 1065)), there is

X y t n e
a-variety-of-environmental-supports(X)

k:=
Y u r v z
calm-but-not-too-motherly-homemaker(u)
referral-for-temporary-economic-aid(r)
intelligent-use-of-nursing-care(v)
accompaniment-to-the-clinic(z)
$Y=u\oplus r\oplus v\oplus z$
X=Y

Parenthetical(k)
$t<n$
$e\subseteq t$
further-development-of-predictable-pathological-mechanisms(y)
e: combine-to-prevent(X,y)

Figure 5: *SDRS for (7)*

a *contrast* between the first text-clause and the second text-clause. It is clear that we run into the problem of under-determination once again. The proposed SDRS is displayed in Figure 7, where the sign $\sum$ is employed as a means to build the sets of leaders (*those who lead*) and followers (*those who follow*).

(10) Those who lead must be considerate; those who follow must be responsive.

In example (11a) (taken from (Quirk et al. 1972: 1068)), a text-phrase constitutes a DRS of its own and serves as an *explanation*. Apart from this, both (11a) and (11b) (also taken from (Quirk et al. 1972: 1068)) can be straightforwardly processed in our given framework. The relevant SDRSs would be as shown in Figure 8.

(11a) There remained one thing he desired above all else: a country cottage.

(11b) In one respect, government policy has been firmly decided; there will be no conscription.

Finally, the reader is invited to consider (12) (taken from (Levinson 1985: 134)):

$$
\begin{array}{|l|}
\hline
\qquad\quad \mathrm{X\ Y\ W} \\
\text{problems(X)} \\
k{:=}\begin{array}{|l|}
\hline
\qquad \mathrm{x\ y\ Z} \\
\text{unemployment(x)} \\
\text{inflation(y)} \\
\mathrm{Z}=\mathrm{x}\oplus\mathrm{y} \\
\mathrm{X}=\mathrm{Z} \\
\hline
\end{array} \\
\text{Parenthetical}(k) \\
\text{economists(Y)} \\
\text{perplex(X,Y)} \\
\text{the-public(W)} \\
\text{mystify(X,W)} \\
\hline
\end{array}
$$

Figure 6: *SDRS for (8)*

(12a) Margaret and Gregory met in 1932, falling in love in a fever of conversation and theory-building on the shores of Sepik River in New Guinea, where Margaret had come to work with Reo Fortune, her second husband.

(12b) Margaret and Gregory met in 1932, falling in love in a fever of conversation and theory-building on the shores of Sepik River in New Guinea. Margaret had come there to work with Reo Fortune, her second husband.

(12c) Margaret and Gregory met in 1932, falling in love in a fever of conversation and theory-building on the shores of Sepik River in New Guinea; Margaret had come there to work with Reo Fortune, her second husband.

In (12a), there is considerable irony which is taken out in (12b), and restored to some degree in (12c). Paragraphs (12b) and (12c) have different interpretations, which are construed in the SDRSs in Figure 9.

Conclusion

Information cues provided by punctuation marks should be valuable to NLP systems if captured carefully and adequately. Our ongoing work is

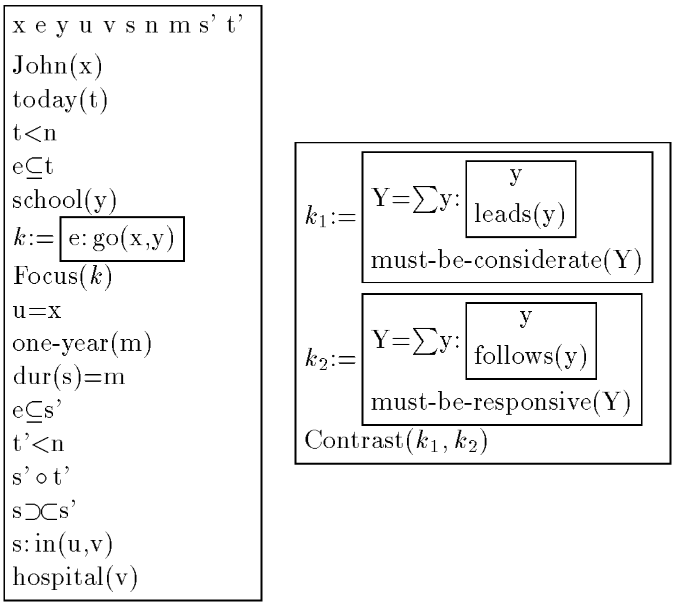

Figure 7: *SDRSs for (9) and (10)*

directed towards classifying the (especially semantic) uses of punctuation with respect to various available resources such as the Wall Street Journal (ACL/DCI 1991) and the SUSANNE corpus (Sampson 1995). At the same time, keeping the interactions with syntax in mind, a system of semantic rules that takes into account the characterization of punctuation marks is being written, extending earlier works by Briscoe (Briscoe 1994) and Lee (Lee 1995). Alvey Natural Language Tools Grammar (Grover et al. 1993), a GPSG-style unification grammar with an event-based and unscoped compositional semantics expressed in λ-calculus, is being used in this endeavor. A computational framework for extracting the informational cues from the actual punctuation practice is planned to be the concrete outcome of this work.

Acknowledgments

The first author is grateful to the Scientific and Technical Research Council of Turkey (program code: BAYG/NATO-A2) for financial aid, and Dr. Ted

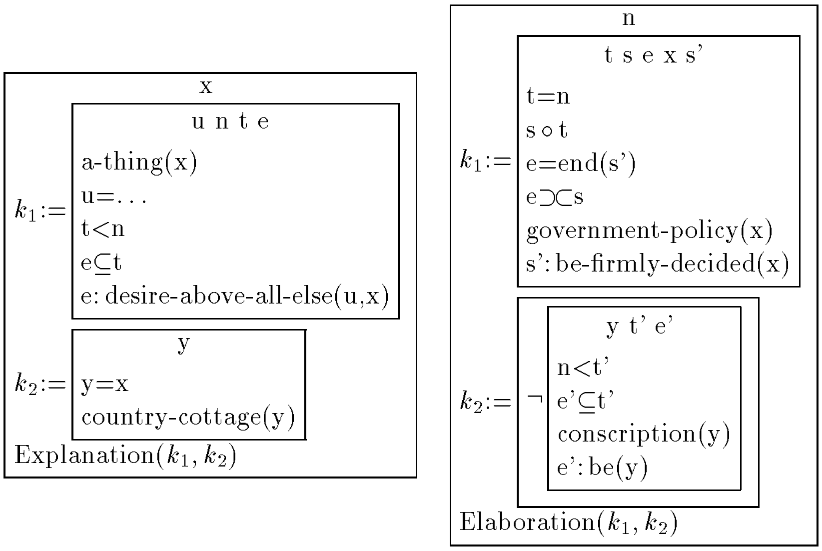

Figure 8: *SDRSs for (11a) and (11b)*

Briscoe for his willingness to accommodate her during a visit in Fall '96 to the Computer Lab., Cambridge University, Cambridge, UK. Finally, our heartfelt thanks to Dr. Carlos Martín-Vide for moral support.

Notes

1. The reader is referred to (Say 1995) for a detailed review of punctuation. Other recent papers of our group which may also be useful include (Bayraktar et al. 1996, Say and Akman 1996b, Say and Akman 1996a).
2. This definition is borrowed from Meyer (Meyer 1986: 80).
3. For a detailed analysis of the role of comma in various types of coordinate compounds, the reader is referred to (Min 1996). A corpus-based study of the semantic functions of comma is reported in (Bayraktar et al. 1996).
4. In this figure and the others in the sequel, the reader may ignore the special symbols appearing in the DRS boxes, when their meaning is not obvious from the context. Thus, only a general understanding of the inner details of the SDRSs is required.

References

x y t n e s u z t' e'
Margaret(x)
Gregory(y)
in 1932(t)
$t < n$
$e \subseteq t$
e: meet(x,y)
$s \circ e$
Sepik River in NG(u)
s: fall-in-love-at(x,y,u)
Reo Fortune(z)
x's second husband(z)
$t' < e$
$e' \subseteq t'$
e': come-to-work-with-at(x,z,u)

x u n e
k_1:= [y t s
Margaret(x)
Gregory(y)
in 1932(t)
$t < n$
$e \subseteq t$
e: meet(x,y)
$s \circ e$
Sepik River in NG(u)
s: fall-in-love-at(x,y,u)]
k_2:= [z t' e'
Reo Fortune(z)
x's second husband(z)
$t' < e$
$e' \subseteq t'$
e': come-to-work-with-at(x,z,u)]
Contrast(k_1, k_2)

Figure 9: *DRS and SDRS for (12b) and (12c)*

ACL/DCI. 1991. Association for Computational Linguistics Data Collection Initiative, CD-ROM 1. Information available from Linguistic Data Consortium on the WWW: http://www.ldc.upenn.edu

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Bayraktar, Murat, Varol Akman, and Bilge Say. 1996. "Analysis of English Punctuation: The Special Case of Comma". Manuscript, Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey. Submitted for publication.

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

Briscoe, Ted. 1996. "The Syntax and Semantics of Punctuation and Its Use in Interpretation". In *Punctuation in Computational Linguistics*, UCSC, Santa Cruz, CA. SIGPARSE 1996 (Post Conference Workshop of ACL96), 1–8. Available from Human Communication Research Center (University of Edinburgh) on the WWW: http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

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

Ehrlich, Eugene H. 1992. *Schaum's Outline of Theory and Problems of Punctuation, Capitalization, and Spelling*. Schaum's Outline Series. New York, NY: McGraw-Hill Book Co.

Grover, Claire, John Carroll, and Ted Briscoe. 1993. "The Alvey Natural Language Tools Grammar". Technical Report 284, Computer Lab., Cambridge University, Cambridge, UK.

Jones, Bernard. 1997. "What's the Point? A (Computational) Theory of Punctuation". Ph. D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Dordrecht, Netherlands: Kluwer Academic Publishers.

Kettunen, Kimmo. 1996. "Low-Level Typographical Spell Checking: A Proposal". *Computers and the Humanities* 30(1): 77–84.

Lee, Sherman. 1995. "A Syntax and Semantics for Text Grammar". Master's thesis, Engineering Dept., Cambridge University, Cambridge, UK.

Levinson, Joan Persily. 1985. "Punctuation and the Orthographic Sentence: A Linguistic Analysis". Ph. D. thesis, City University of New York, New York, NY.

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

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

Meyer, Charles F. 1986. "Punctuation Practice in the Brown Corpus". *ICAME Newsletter*, 80–95.

Meyer, Charles F. 1987. *A Linguistic Study of American Punctuation.* New York, NY: Peter Lang Publishing Co.

Min, Young-Gie. 1996. "Role of Punctuation in Disambiguation of Coordinate Compounds". In *Punctuation in Computational Linguistics*, UCSC, Santa Cruz, CA. SIGPARSE 1996 (Post Conference Workshop of ACL96), 33–40.

Available from Human Communication Research Center (University of Edinburgh) on the WWW: http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation.* Number 18 in CSLI Lecture Notes. Center for the Study of Language and Information, Stanford, CA: CSLI Publications.

Osborne, Miles. 1996. "Can Punctuation Help Learning?". In S. Wermter, E. Riloff, and G. Scheler (eds), *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, Lecture Notes in Artificial Intelligence, Number 1040. Berlin: Springer-Verlag, 399–412.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1972. *A Grammar of Contemporary English.* Harlow, Essex, UK: Longman.

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

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

Say, Bilge and Akman, Varol. 1996a. "Current Approaches to Punctuation in Computational Linguistics". Manuscript, Dept. of Computer Engineering and Information Science, Bilkent University, Ankara, Turkey. Submitted for publication.

Say, Bilge and Akman, Varol. 1996b. "Information-Based Aspects of Punctuation". In *Punctuation in Computational Linguistics*, UCSC, Santa Cruz, CA. SIGPARSE 1996 (Post Conference Workshop of ACL96), 49–56. Available from Human Communication Research Center (University of Edinburgh) on the WWW: http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html.

Vallduví, Enric. 1992. *The Informational Component.* New York, NY: Garland Publishing.

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