AN INFORMATION-BASED APPROACH TO PUNCTUATION

A DISSERTATION SUBMITTED TO
THE DEPARTMENT OF
COMPUTER ENGINEERING AND INFORMATION SCIENCE
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BİLKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

by
Bilge Say
November 1998

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Varol Akman (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Asst. Prof. İlyas Çiçekli

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Asst. Prof. David Davenport

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Assoc. Prof. Haldun Özaktaş

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Asst. Prof. Ümit Deniz Turan

Approved by the Institute of Engineering and Science:

Prof. Mehmet Baray
Director of the Institute of Engineering and Science

ABSTRACT

AN INFORMATION-BASED APPROACH TO PUNCTUATION

Bilge Say
Ph.D. in Computer Engineering and Information Science
Supervisor: Prof. Varol Akman
November 1998

Punctuation marks have special importance in bringing out the meaning of a text. Geoffrey Nunberg's 1990 monograph bridged the gap between descriptive treatments of punctuation and prescriptive accounts, by spelling out the features of a text-grammar for the orthographic sentence. His research inspired most of the recent work concentrating on punctuation marks in Natural Language Processing (NLP). Several grammars incorporating punctuation were then shown to reduce failures and ambiguities in parsing. Nunberg's approach to punctuation (and other formatting devices) was partially incorporated into natural language generation systems. However, little has been done concerning how punctuation marks bring semantic and discourse cues to the text and whether these can be exploited computationally.

The aim of this thesis is to analyse the semantic and discourse aspects of punctuation marks, within the framework of Hans Kamp and Uwe Reyle's Discourse Representation Theory (DRT) (and its extension by Nicholas Asher, Segmented Discourse Representation Theory (SDRT)), drawing implications for NLP systems. The method used is the extraction of patterns for four common punctuation marks (dashes, semicolons, colons, and parentheses) from corpora, followed by formal modeling and a modest computational prototype. Our observations and results have revealed interesting occurrences of linguistic phenomena, such as anaphora resolution and presupposition, in conjunction with punctuation marks. Within the framework of SDRT such occurrences are then tied with the overall discourse structure. The proposed model can be taken as a template for NLP software developers for making use of the punctuation marks more effectively. Overall, the thesis describes the contribution of punctuation at the orthographic sentence level to the information passed on to the reader of a text.

Keywords: Punctuation, Discourse, (Segmented) Discourse Representation Theory [(S)DRT], Information Structure, Corpora, Natural Language Processing (NLP)

ÖZET

NOKTALAMAYA ENFORMASYON TEMELLİ BİR YAKLAŞIM

Bilge Say
Bilgisayar ve Enformatik Mühendisliği Doktora
Tez Yöneticisi: Prof. Dr. Varol Akman
Kasım 1998

Yazılı dilin anlamsal ifadesinde noktalama işaretleri özel bir önem taşır. Geoffrey Nunberg'in yazılı cümlede noktalama işaretlerinin oluşturduğu metin grameri üzerine 1990 tarihli kitabı bu konudaki betimleyici ve buyurucu yaklaşımları birleştirmiştir. Bu yapıt yakın geçmişte Doğal Dil İşleme (DDİ) alanında noktalama işaretlerine yaklaşımların çoğuna esin kaynağı olmuştur. Daha sonra geliştirilen sözdizimsel ayrıştırıcılar çözümleme hata ve belirsizliklerinin noktalama işaretlerinin göz önüne alınmasıyla azaldığını göstermiştir. Keza Nunberg'in noktalama işaretlerinin (ve metin düzenleme araçlarının) sunumuna getirdiği yaklaşım doğal dil üretme dizgeleri tarafından değerlendirilmiştir. Ancak noktalama işaretlerinin anlamsal ve söylemsel etkileri ve bunların hesapsal kullanımı hakkında çok az çalışma yapılmıştır.

Bu tezin amacı noktalama işaretlerinin anlamsal ve söylemsel yönlerini Hans Kamp ve Uwe Reyle'nin Söylem Gösterim Kuramını (SGK) (ve Nicholas Asher'in bunun üzerine geliştirdiği Bölümlü Söylem Gösterim Kuramını (BSGK)) kullanarak incelemek ve DDİ dizgeleri için gerekli sonuçları çıkarmaktır. Uygulanan yöntem elektronik metinlerden dört yaygın noktalama işareti (uzun tire, noktalı virgül, iki nokta üstüste ve parentez) ile ilgili örüntüleri çıkararak, biçimsel bir model ve bilgisayarda küçük bir uygulama elde etmek olarak özetlenebilir. Gözlem ve sonuçlarımız anafora çözümleme ve varsayım gibi dilbilimsel olguların noktalama işaretleri ile ilgisi hakkında ilginç bağlar ortaya çıkarmıştır. BSGK çerçevesinde bu örneklemeler genel söylem yapısına bağlanmıştır. Önerilen model DDİ için yazılım geliştirenlerin noktalama işaretlerini daha etkili kullanabilmesi için bir şablon olarak alınabilir. Tez genelde noktalamanın yazılı metin aracılığıyla okuyucuya aktarılan enformasyona yaptığı katkıyı betimlemektedir.

Anahtar Sözcükler: Noktalama, Söylem, (Bölümlü) Söylem Gösterim Kuramı [(B)SGK], Enformasyon Yapısı, Külliyat, Doğal Dil İşleme (DDİ)

ACKNOWLEDGMENTS

To Füsun Aktan, to the vivacious
and tender person she was.

I wish to express my deepest gratitude to my supervisor, Dr. Varol Akman for encouragement, patience, and support both for the subject matter of this thesis and for leading me through the intricacies of this period of preparation for an academic life.

Several people helped the thesis become a reality. I would like to thank my committee members for their suggestions and comments. I would like to thank Dr. Ted Briscoe for letting me visit their lab at Cambridge University, for allowing me to use their resources, and for comments. I benefited from the comments of various researchers as I presented my work on several occasions. Others directed me to useful references and contacts. I would especially like to thank Drs. David Beaver, Cem Bozşahin, Pierre Flener, Nancy Ide, Bernard Jones, Geoffrey Nunberg, Gwen Robinson, Şükriye Ruhi, Jerry Seligman, Candy Sidner, Carlos Martín-Vide, and several anonymous reviewers. Thanks to Dilek Hakkani Tür and Pınar Saygın for proofreading. Obviously, I am solely responsible for the contents of the thesis.

Bilkent University funded me throughout my Ph.D. studies; TÜBİTAK and AAAI gave partial support for international activities. I would like to thank NATO TU-LANGUAGE project and its principal investigator Dr. Kemal Oflazer for allowing me to use their resources.

Finally, to family and friends, I owe lots of THANKS! To Okyay, who I know would never marry another person about to start her Ph.D. studies and with whom I look forward to discovering life again. To my father, for being my unofficial advisor and to my mother, for her serene support. To my "family-in-law" for their encouragement and prayers. And to those who are in warm corners of my heart for being best friends.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

$*$	any category
$+$	repetition of the referred entity one or more times
$\mid$	choice between divided items
$\Rightarrow$	condition in DRT
$\Downarrow$	discourse domination in SDRT
$<$	temporal precedence in DRT
$\subseteq$	temporal or nominal inclusion in DRT
$\circ$	temporal overlap in DRT
$\mid \ldots \mid$	number of elements in a set
$\oplus$	summation of entities
eos	end of sentence mark (period, exclamation or question mark)
ADJP	Adjectival Phrase
ADVP	Adverbial Phrase
ANLT	Alvey Natural Language Tools
BC	Brown Corpus
BNC	British National Corpus
CC	Coordinating Conjunction
DET	Determiner
DRS	Discourse Representation Structure
DRT	Discourse Representation Theory
GDE	Grammar Development Environment
GPSG	Generalized Phrase Structure Grammar
ID	Immediate Dominance

N	Noun
NP	Noun Phrase
NLP	Natural Language Processing
NLG	Natural Language Generation
PP	Prepositional Phrase
PS	Phrase Structure
RST	Rhetorical Structure Theory
S	Sentence
SBAR	Clause Introduced by Subordinating Conjunction
SDRS	Segmented DRS
SDRT	Segmented DRT
SEC	Spoken English Corpus
VP	Verb Phrase
WSJ	Wall Street Journal Corpus

Chapter 1

Introduction

1.1 Motivation

Punctuation marks, the symbols that assist the understanding of written text, have usually been regarded as conventions, thus as being outside the domain of pure linguistics. However, even as conventions, they have interacted with the written language for centuries. Nowadays, they no longer have a singular function such as helping reading aloud. Several researchers [Nunberg, 1990, Meyer, 1987, Robinson, 1997] observed the need for a linguistic study of punctuation. Most notably and recently Nunberg [1990] hypothesized that punctuated sentences form a *text-grammar* of their own as opposed to a (lexical) grammar in the linguistic sense. He also sketched such a grammar syntactically. A natural question is whether semantic or discourse contributions of punctuation can be characterized linguistically and how available corpora can be used computationally to extract these. This is the primary motivation underlying this thesis.

It might be useful here to give an idea of the kind of effects we have in mind, with sentences from well-known corpora:

(1.1) **(WSJ)**[1] At one point, almost all the shares in the 20-stock Major Market Index, which mimics the industrial average, were sharply higher.

(1.2) **(BC)** It is a killer sub—that is, a hunter of enemy subs.

[1] See the List of Symbols and Abbreviations

(1.3) **(WSJ)** Unless other rules are changed, the devaluation could cause difficulties for the people it is primarily meant to help: Soviets who travel abroad.

In (1.1), the commas that surround the which-clause signal that the clause is a non-restrictive one. In other words, the meaning of the clause would be interpreted differently if it were not giving extra information about a certain index but distinguishing it from a number of indices having the same name. In (1.2), the reader is presumed to have a knowledge of what a *killer sub* is. The dash acts as a cue for those who lack that knowledge. Technically speaking, this is a certain way of accommodating a presupposition (see Chapters 3 and 4). In (1.3), dislocating *Soviets who travel abroad* by means of a colon has enabled placing emphasis on this constituent. (This may further have effects on how to resolve anaphora; cf. Section 3.5.)

Every written sentence, on the average, contains four punctuation marks including the period (finding on SEC [Jones, 1997]). Over 50% of the sentences in a corpus contain some punctuation mark other than a period (finding on SUSANNE [Briscoe, 1996]). Therefore, even the mere frequency of observed marks makes punctuation a viable and worthy subject for investigation. Previous studies [Briscoe, 1996, Jones, 1997] have already shown that parsing failures and ambiguities decrease when syntactic patterns of punctuation are taken into account. The question that acts as a computational catalyst for this thesis is whether one can also capture the semantic and discourse effects of punctuation in NLP modules in a principled way.

1.2 Objectives

The objectives of this thesis can be listed as follows:

- To conduct a linguistic study of the semantic and discourse effects of punctuation on various corpora of written English.

- To link the findings with the structure of the sentences.

- To model the semantic and discourse effects of punctuation within a contemporary theory (i.e., (S)DRT) that takes context into account.

1.3 Methods

The objects of this study are punctuated sentences in English texts (primarily by native, adult authors and in non-literary genres). Attention is focused on four commonly used marks that act at (lexical) phrase, clause, or sentence levels: dash [—], semicolon [;], colon [:], and parentheses[()]. Comma , the most versatile of marks, is not studied in detail in this thesis except for some examples of its semantic effects to be modeled in Chapter 4. After all, it has been investigated by us in reasonable depth [Bayraktar *et al.*, 1998] and interesting results have been obtained in linking syntactic patterns to functional usage.

Techniques from corpus linguistics and computational linguistics are used as well as formal semantic modeling. The corpus-based approach involves computer-based scripts (pattern matchers) to extract and classify relevant data from corpora as well as direct observations on the sentences extracted. Formal semantic modeling is performed via a current and respected semantic theory, viz. (S)DRT [Kamp and Reyle, 1993, Asher, 1993].

1.4 Outline

In the next chapter, existing works on punctuation are summarised and evaluated, accompanied by a historical perspective. In Chapter 3, a linguistic characterization of the semantic and discourse effects of punctuation is given via a corpus-based study on our four selected marks. In Chapter 4, the linguistic characterizations are described within the formal semantic framework of SDRT. The reader who would like to see an assessment of the contributions of this thesis can refer to Chapter 5 which summarises them along with shortcomings, limitations, and suggestions for future research. Appendices A and B provide background material on discourse relations and SDRT. Appendix C offers a summary of computational work.

Chapter 2

Current Approaches to Punctuation in (Computational) Linguistics

2.1 Introduction

Punctuation has not been studied much by linguists apart from a prescriptive standpoint until the eighties. Similarly, most NLP systems did not take punctuation marks into account except for the period. However, there have been recent works in linguistics (computational, corpus, and applied), giving a descriptive treatment of the role of punctuation. Furthermore, various NLP systems have started to make use of the syntactic cues provided by the punctuation marks. This chapter presents the current state of incorporation of punctuation marks into NLP systems [Say and Akman, 1997]. Concentration is on punctuation in English; there exists some work on punctuation in other languages [Akram and Saadeddin, 1987, Simard, 1996, Twine, 1984].

Throughout the chapter, punctuation marks are taken to be not only the standard ones such as the comma, colon, dash, etc., but also the more graphical devices such as paragraphs, lists, emphases (e.g., italics), etc. Essentially, any feature that can shape orthographically written text into comprehensible units [Robinson, 1988, p. 75] is within our coverage.

Punctuation is traditionally considered different from other language elements [Pullum, 1991]: It is due to invention, not evolution along with species. It constitutes a learned system in which mastery is not common. Moreover, it seems (according to

Pullum) more natural, compared to other elements of written text, to take a prescriptive approach towards punctuation. Even if there are elements of truth in Pullum's observations, conventional systems such as punctuation tend to have patterns of their own at least in writing by adult, native writers of a language (English). Therefore, descriptive and formal treatments of such patterns with possible uses in NLP are worth the effort.

In general, one can come up with different classifications of punctuation marks. One classification is according to whether a text is punctuated for the ear or the eye. Elocutionary punctuation emphasizes the rendering of the written text as close to the spoken word as possible by way of pauses, etc. Logical (or syntactic) punctuation emphasizes the structuring of the sentence.

Another classification of punctuation can be made according to the units a mark acts on, cf. Jones [1997, pp. 4–8]. Marks that occur between lexical items (e.g., comma, semicolon, etc.) are called *inter-lexical* marks. In this thesis, the term *structural* marks, due to Meyer [1986, p. 80] will be preferred. Marks that occur (usually) within words (e.g., hyphen, apostrophe) are called *sub-lexical*. Sub-lexical marks are better defined and documented than other kinds of marks in how they change the meaning of a word. Other orthographic processes that characterize text (e.g., paragraphing, underlining, etc.) are called *super-lexical* (or *text*) punctuation [Pascual and Virbel, 1996]. Structural (inter-lexical) punctuation will be the subject matter of this thesis.

A linguistic and computational survey of punctuation will be given in the remainder of this chapter. Section 2 gives a perspective on the history of punctuation and its place in writing today. In Section 3, current linguistic studies are presented, excluding the computational ones. In Section 4, relevant NLP works on the relationship of syntax and punctuation are evaluated. In Section 5, semantic, intonational, and discourse implications of punctuation are discussed.

2.2 Punctuation and Written Language

According to Parkes [1993], the development of punctuation took place in several stages paired up with the development of the written medium. Each stage's reader group required different demands to be satisfied, thus affecting the marks and their functions. In

Classical Latin writing, education was directed at preparing students for effective public speaking [Parkes, 1993, p. 5]. Authors often dictated their writing to the scribes or the scribes copied from manuscripts. Because the scribes usually did not understand the material they were copying, the usage of punctuation was very much varied [Robinson, 1989, p. 73]. Spaces between lexical words did not become customary until the tenth century [Levinson, 1985, p. 23]. As opposed to punctuating for oral readers, some grammarians saw writing as a means for silently conveying meaning to the reader [Parkes, 1993, p. 21]. During the eighth century, the Irish devised new graphic conventions in the written text and later passed those conventions onto the Anglo-Saxons [Parkes, 1993, p. 23]. From the 12th century onwards, a general inventory of punctuation marks was designed but, since even two scribes copying the same manuscript employed different marks, there was no standardisation [Parkes, 1993, p. 69].

Rhetorically organized speech shaped the text according to the principles of spoken art before the medieval era [Robinson, 1988, p. 94]. Then, writing started to go beyond the boundaries of the monasteries. As it was gradually used for secular purposes, economy and speed in reading became more important [Levinson, 1985, p. 38]. Writers started to use punctuation to bring out the relationships between the grammatical constituents of the sentence. In particular, during the 14th to 16th centuries, the humanists wanted their texts to be persuasive. Thus, they adopted a larger set of punctuation marks to disambiguate the logical structure of sentences. New marks corresponding to today's parentheses, semicolon, and exclamation mark were devised in the 15th and 16th centuries. From the 16th century onwards, with the widespread usage of printing, a gradual standardisation emerged. Types and fonts were precut and sold to printers; so the available repertory of marks was no longer personalised by the scribes. Also, before printing, the destination of the manuscript being prepared (e.g., a specific monastery or library) was mostly known beforehand. After printing became the norm, this connection between the publisher and the client was broken; there was now a greater pressure for general understandability of the text. The orthographic sentence became the fundamental unit presented to the reader [Levinson, 1985, p. 157]. Rhetorical question marks, apostrophes, quotation marks, and italics (yielding emphasis) emerged after the 16th century.

In the last quarter of the 16th century, “writing became more purposeful, direct and fact-oriented.” [Robinson, 1990a, p. 113]. This tipped the balance in favour of logical (as opposed to elocutionary) punctuation. Sentences became considerably long. On 19 December 1700, a letter by a Mr. Prior to a Mr. Talbott was several pages long; yet, it consisted of a single sentence [Robinson, 1990b, p. 97]. In the 18th and 19th centuries, assorted books and articles were written on English punctuation [Robinson, 1990b, p. 102]. Publishers established a simpler, cost-effective set of principles for punctuation [Robinson, 1992, p. 113]. Punctuation for the rhetorical and logical structure of the text became so widespread that the early 20th century novelists frequently used punctuation to create the so-called “stream of consciousness” effects [Parkes, 1993, p. 87].

Towards mid 20th century, as radios and telephones became widespread, a shorter and sharper language of factual and scientific style became more valuable [Robinson, 1996, p. 75]. With the addition of TV and computers there is nowadays even more emphasis on keeping the written text simple, quick, and close to “sensorial immediacy” [Robinson, 1997, p. 130]. Between 1936 and 1996, the average sentence length in best-selling books (in the US) decreased by two fifths, while the amount of dialogue increased by a third. As for punctuation, its frequency of use has dropped nearly 50% (except for the period) [Robinson, 1997, p. 127]. However, works emphasising the usage of punctuation marks in modern texts [Jones, 1995, Meyer, 1986] authoritatively state that punctuation is still an integral part of the written language. A study done on nine different corpora of current English shows that a typical English sentence is likely to contain two to five punctuation symbols, and a punctuation mark of some variety is likely to be encountered on average every fourth to seventh word [Jones, 1997, p. 87].

Punctuation marks have also been studied as a system of signs from a semiotical point of view [Harris, 1995]. Harris does not regard a writing system as being simply projected from speech. Rather, written signs are analyzed according to their related types of activity (forming, processing, and interpretation) [Harris, 1995, p. 60]. Writing uses spatial relations and, thus, is different from speech. In understanding forms of punctuation such as tabular writing, which has no counterpart in spoken language, the internal syntagmatics (i.e., “the disposition of written forms relative to each other”

[Harris, 1995, p. 121]) becomes crucial.

2.3 Linguistic Work on Punctuation

Style guides and grammar books [Ehrlich, 1992, McDermott, 1990, Partridge, 1953] usually offer a prescriptive account of punctuation. As for the applied linguistic arena, there are mostly works relating to learnability. Scholes and Willis [1990] recite an experiment where university students, when asked to read a text aloud, interpreted punctuation marks as elocutionary even when the marks had other (semantic) effects. Smith [1986] describes another experiment to determine whether a graphical instruction environment is better liked by students learning punctuation. A recent project tackles the question of how young children understand the nature and use of (English) punctuation; the aim is to find effective ways of teaching punctuation [Hall and Robinson, 1996].

The first up-to-date descriptive treatment of punctuation as a system is Meyer's book [1987]. He concentrates on the American usage of *structural* punctuation marks, that is, 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) [Meyer, 1986, p. 80]:

> This study focused exclusively on "structural punctuation": periods, question marks, exclamation marks, commas, dashes, semicolons, colons, and parentheses. It did not deal with paragraph indentations (or separation) or apostrophes and hyphens, nor did it focus on brackets, ellipsis dots, quotation marks, and underlining, or the use of commas and colons in dates, times, etc. These are marks of punctuation whose uses have been fairly rigidly conventionalised by style manuals.

While structural marks are a good working category to distinguish from text punctuation (such as paragraphs, font changes, lists), the definition given is not exactly correct. For instance, it is obvious that parentheses occasionally do work on units larger than sentences. In fact, this is one of the reasons for Dale's [1991a] call for a theory of discourse (and discourse uses of punctuation) spanning the sentence boundary.

Meyer uses 12 samples, approximately 2,000 words each, from the BC [Francis and Kučera, 1982]. He classifies and exemplifies the functions of punctuation, and how those functions are realised. Distinguishing between the functions of marks and their realisations is a point he stresses to be usually missing from the prescriptive work. Functions basically help the reader understand efficiently and easily, emphasise a construction, or vary the rhythm of the text. He groups their realisation into two categories: marks that separate (such as periods, colons) and marks that enclose (such as dashes, parentheses). He then gives a detailed account of the boundaries that punctuation marks work on: syntactic (clauses, phrases, or words), prosodic (pauses, tone units, and changes in stress and pitch), and semantic (questions, modifiers, etc.). He notes that punctuation usually overdetermines—determines more than one kind of boundary—but that it usually favours one more than the other.

Meyer's work is the first of its kind in synthesising a linguistic account of punctuation from corpus data. His book is valuable in comparing what style manuals prescribe and what actually happens. However, the size of his samples is too small (compared to what is available nowadays). His linguistic analysis, while generally complete, amounts to observations rather than generalizations.

Levinson [1985] offers a historical perspective on the development of punctuation. She sees two serious flaws in recent works. One is that "Punctuation marks syntax". The other is that "The fundamental entity which determines punctuation is the sentence". She observes a potential circularity in trying to establish rules according to the distribution of punctuation. The rules require a prior notion of sentence. Yet a clear definition of sentencehood must be based on punctuation, namely capital letters and the period! She proposes a way out of this circularity by separating the grammatical sentence from the orthographic one. She claims that relating punctuation to syntax may stem from the fact that it is easier to do so; other linguistic features such as intonation contours or semantic concepts would make it more difficult. She proposes to view the orthographic sentence as an informational grouping based on (but distinct from) syntactic structure and specified by the rules of punctuation (not grammar). She defines *informational grouping* as putting, within the limits of the orthographic sentence, the linguistic units in the right order according to their informational links. She goes on to describe the linguistic units she uses for this purpose (i.e., proper clause structures and sentence partials) and gives a classification of the actual grouping. Sentence partials like adverbial clauses and tenseless verb phrases, as Levinson sees them, do not classify as proper clauses. In attaching sentence partials to proper clauses and to other sentence partials, a signal of attachment (an informational link) is required. Various devices can act as such a signal, e.g., conjunctions and phrase ordering. Punctuation is also one of them. Consider the following examples, [Levinson, 1985, p. 130] with different kinds of attachment:

(2.1) a. He was happy to find his book.

b. He was happy because he found his book.

c. He was happy. He found his book.

In (2.1c), a limit to the informational group "He was happy" has to be put by means of a period. Where (how) a sentence partial is attached (presented) gives rise to different information groupings [Levinson, 1985, p. 134].

The book on which the majority of the studies reviewed in the next section are based is [Nunberg, 1990]. Nunberg attributes the negligence of punctuation in the linguistic community to its being relatively new as well as its being perceived as prescriptive and a reflection of intonation. He explains that the origin of punctuation was the transcription of intonation but then the two diverged; now punctuation is a linguistic system in its own right. He describes a *text-grammar* as the collection of rules that explains the distribution of explicitly marked categories such as paragraph, sentence, or parentheticals. He usually excludes semantic or pragmatic relations of coherence and the like from his definition of text-grammar, as these depend on context.

Nunberg constructs his text-grammar so that it accounts for punctuation marks between text-categories (text-clauses, text-adjuncts, or text-phrases) which are themselves dealt with by the lexical grammar. He proposes various rules for English to handle the interactions between various marks. One such rule, for example, is the point absorption rule, which among other things dictates that a period will absorb a comma when they are adjacent.

Two reviewers of Nunberg [Humphreys, 1993, Sampson, 1992] acknowledge his work positively. Sampson observes several counter-examples to Nunberg's rules though, remarking that they are not adequately based on empirical data. Switching between single and double quotations is not uniformly distinguished between American and British practices. Brackets or colon-expansions can be nested as opposed to Nunberg's suggestion (a point also noted by Jones [1997]). These kinds of stylistic choice clearly make the task of establishing a set of tidy, empirical rules for punctuation harder.

Nunberg's way of deciphering punctuation as a linguistic subsystem separate from but related with (lexical) grammar has been a starting point for other research (see also [Jones, 1996a]). When a unified theory of punctuation is born, it may not be like what Nunberg has suggested in particulars. But it has to account for the issues first raised and studied by him [Nunberg, 1997].

In all, most of these works recognise the information-providing function of punctuation marks. However, they do not attempt to propose a formal account, apart from Nunberg's work, which eloquently covers the syntactic and presentational aspects of punctuation.

2.4 Computational Work on Punctuation

Computational linguists have worked on the recognition of sentence boundaries for part-of-speech tagging and sentence alignment in bilingual corpora. Palmer and Hearst [1994] use a neural network with part-of-speech probabilities to label sentence boundaries. Reynar and Ratnaparkhi [1997] use a maximum entropy model (for training) that requires little prior information to detect valid boundaries.

Garside and his colleagues [1987] describe a research programme undertaken between 1976-1986. Their aim was to base NLP on the probabilistic analysis of a large corpus. In describing the tagging subsystem, they take punctuation marks (tagged to delimit ambiguity) into account. A related project on "automatic intonation assignment" aims to produce a prosodic transcription from written versions of punctuated, spoken texts.

Also worth mentioning is the SUSANNE analytic scheme [Sampson, 1995], a notation for indicating the structural (grammatical) properties of texts taken from the Brown Corpus [Francis and Kučera, 1982]. SUSANNE is a comprehensive, consistent, and theory-neutral notation that will be of use to researchers working on corpora. Punctuation marks have their own tags and act as leaf nodes in a SUSANNE parse tree. Various ambiguities as to where to attach them within the parse tree are worked out.

Jones [1995, 1997] has made a computational analysis of the structural punctuation marks on various corpora, including the Guardian newspaper (12 million words), the Leverhulme Corpus (a corpus of student essays comprising 356,000 words), the WSJ (184,000 words), and articles extracted from the Usenet. He computed the percentages of various marks and compared the complexity and genre of the texts with the frequency of the marks.

A natural language understanding system that takes punctuation into account is the Constraint Grammar developed by Karlsson and his colleagues [1994]. This is an effort for morphological and syntactic parsing of language-independent, unrestricted text. Karlsson *et al.* combine a grammar-based approach with optional heuristics, when the former fails. The emphasis is on discarding improper alternatives by means of constraints, which are rules for disambiguation. The aim is to simplify parsing through the use of typographical features such as punctuation, case (of letters), and mark-up (of texts). They treat all sentence delimiters plus non-letter and non-digit characters as specially-marked, individual words which may have features and be referred to by constraints. In this way, punctuation marks are used to detect clause boundaries or lists of similar categories; they are also used to implement heuristics as in the observation that certain punctuation marks (such as dashes that are to the left of a finite verb) dramatically decrease the probability of the preceding word being a subject.

Jones [1994a, 1994b, 1996b, 1997] describes parsing-related work based mainly on Nunberg's framework, using a feature-based tag grammar. He refrains from using a two-level (lexical and text) grammar as advocated by Nunberg on the grounds that interactions between the levels make the grammar unnecessarily complex [Jones, 1994b]. For Nunberg, the lexical expressions must have information about their neighbouring syntactic categories so that the text grammar can draw proper conclusions. Jones instead modifies an existing grammar for English by introducing a notion called *stoppedness* for a category that describes the punctuation mark (if any) following it. The rules cater for the optionality of certain marks and the absorption rules (e.g., a period absorbing an adjacent comma) through *stop* values. Testing his grammar on the SEC [Taylor and Knowles, 1988], which includes rich punctuation, he concludes that the number of parses is significantly reduced. He also introduces a measure of complexity (of a sentence) in terms of punctuation; there is a direct relationship between the number of parses a given sentence has and the average number of words residing between two punctuation marks in it. Jones revises his implementation methodology in later works [Jones, 1996b, Jones, 1996c, Jones, 1997]. For instance, discarding stoppedness ensures better modularity. He draws 79 generalized syntactic punctuation rules (regarding colon, semicolon, dash, comma and period) from nine corpora. His revised grammar produces similar (or even slightly better[1]) results compared to Briscoe and Carroll [1995].

Briscoe and Carroll [1994, 1995] build a text-grammar as advocated by Nunberg, by tokenising punctuation marks separately from words. Punctuation is seen as useful for not only breaking the text into suitable units for parsing but also for resolving structural ambiguity. They build a punctuation grammar for capturing text-sentential constraints described by Nunberg and integrate this grammar into another for part-of-speech analysis. Treating text categories and syntactic categories as overlapping, and dealing with disjoint sets of features in each grammar render the integration to be more modular than the approach taken by Jones. They test the resulting grammar on SEC and SUSANNE and give detailed interpretations of their results [Briscoe and Carroll, 1995]. When about 2,500 in-coverage (covered by the resulting grammar) SUSANNE sentences were stripped off of their punctuation, around 8% of them failed to receive an analysis at all and an average sentence received 38% more parses than before. Lee syntactically and semantically extends the grammar described above [Lee, 1995]. She implements the distinguishing semantics between subordinating and coordinating constructs. Upon testing her grammar on a small corpus, she finds that syntactically all the punctuated sentences have at least one parse whereas 50% of the same sentences do not parse at all when they are left unpunctuated [Briscoe, 1996].

Doran's work concentrates on the role of punctuation in quoted speech [Doran, 1996]. A detailed analysis of the role of comma in various types of coordinated compounds is

[1] An exact comparison is not possible as they use different core grammars and Jones deletes 300 sentences from his data set because they are outside the coverage of his grammar.

given in [Min, 1996]. Shiuan and Ann [1996] report an experiment about separating complex sentences with respect to punctuation and parsing the so-created chunks first. They observe a 21% error reduction in parsing as compared to the performance of their original parser. Osborne [1996] recites an experiment where even a simplified model of punctuation enhanced learning unification-based grammars. Kessler, Nunberg, and Schütze [1997] use punctuation as one of the surface cues for the classification of text into genres. An obvious cue coming from punctuation is the count of occurrences of say, question marks which are indicative of certain genres.

White [1995] investigates how Nunberg's approach to presenting punctuation (and other formatting devices) might be incorporated into NLG systems. He criticises Nunberg's analysis of punctuation presentation rules, giving examples where some options work fine from a parsing point of view but overgenerate from a generation point of view. He then proposes a layered architecture which has three components: syntactic, morphological, and graphical. The components deal with punctuation presentation rules for hierarchy, adjacency, and graphical form, respectively. An implementation of punctuation and formatting-rules has been incorporated into a generation system that produces the final text of a target language according to syntactic, morphological, and lexical constraints [Lavoie and Ranbow, 1997]. Reed and Long [1997] describe a general framework for the generation of natural language arguments. They propose an intention and salience-based way of generating quotations, footnotes, etc.

As can be seen, there is considerable recent work on using punctuation marks (especially for the task of syntactic parsing) and characterising their usage with corpora. As to the systems described [Garside *et al.*, 1987, Karlsson *et al.*, 1994], it is hard to say to what degree they incorporate punctuation. From a parsing point of view, Briscoe and Carroll's [1995] and Jones' [1997] systems are significant. More work on specific marks such as quotations [Doran, 1996] will prove to be valuable. The next question is whether the works cited above cover enough ground to fully characterise punctuation.

2.5 Other Aspects of Punctuation

2.5.1 Discourse and Punctuation

Previous research under this particular heading has consisted mostly of examples and ideas that have not been methodically tested. Consider the following sentences from Nunberg [1990, p. 13]:

(2.2) a. Order your furniture on Monday, take it home on Tuesday.

b. Order your furniture on Monday; take it home on Tuesday.

Nunberg indicates that (2.2a) has a conditional sense whereas (2.2b) is merely a conjunction. Now consider the following, again from Nunberg [1990, p. 31]:

(2.3) 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.

In (2.3a) the spokesman (*He*) announced the decision—that they were forbidden to speak with the chairman directly. In (2.3b) the spokesman reported the decision to the chairman as others were forbidden to speak with the chairman directly. In a less intuitive setting, (2.3b) can also mean that the reason the spokesman announced the decision himself (rather than the chairman) was that they were forbidden to speak with the chairman directly.

The relationship between discourse and punctuation that these examples suggest has also been noted by Dale [1991a, 1991b]. He raises questions about what roles punctuation plays within discourse structure. He points out to the relationship among discourse markers[2], punctuation marks, and graphical markers (such as paragraph breaks or lists). Punctuation marks are not openly linguistic as cue words nor openly layout-oriented such as lists but they at times perform similar functions.

[2]Discourse markers (also known as cue words) [Schiffrin, 1987] are lexical markers aiming to bring to the listener's attention the bond between the next utterance and the current discourse context. Examples include *well, therefore, thus,* etc.

Dale observes that many uses of certain marks (colon, semicolon, dash, parentheses, comma) act as signals of discourse structure usually within the orthographic sentence level. This justifies the need for a discourse theory that should be able to operate below and above the orthographic sentence level. Discourse structure involves a hierarchical structuring of text units according to relationships between them. *Discourse relations* act as glue by indicating implicit relations between those parts so that the content of one part may, for example, elaborate, exemplify, or explain that of another. This idea forms a central part of this thesis and will be reexamined in the sequel.

Dale states that punctuation underdetermines discourse relations in a text since the same marks can be used for different relations. He considers the possibility of taking a syntactic view of punctuation within discourse. This might involve, for example, determining whether one segment serves as a precondition for another without assigning exact discourse relations. He tries preliminaries of both an intentional structure and a coherence structure by respectively using the approach of [Grosz and Sidner, 1986], and the Rhetorical Structure Theory (RST) [Mann and Thompson, 1987]. RST involves characterising discourse (or coherence) relations that hold between arbitrarily long units of text. Relationships are numerous (including elaboration, justification, etc.) and can be applied hierarchically. (See Appendix A for a core subset of RST relations.)

Dale's suggestions are extended in this thesis in a more concrete way in multiple directions: linking syntax and semantic effects, linking discourse effects with discourse relations, and linking the linguistic observations with computational modeling.

2.5.2 Intonation and Punctuation

There is also a parallel between intonation (and the efforts to formalise it) and punctuation. Cruttenden [1986] explains that for many uses of punctuation there is no intonational equivalent. Some exceptional uses usually correlate with the boundaries of a separate intonation group such as a pair of commas in parenthetical use. He claims that the often unnecessary usage of a comma between the subject and the predicate of a clause occurs from such a coincidence. Bolinger [1989], on the other hand, has investigated the relationship of intonation to discourse and grammar. He thinks that intonation and grammar are pragmatically (but not linguistically) interdependent, but this interdependence is not a strict one. He produces cases where punctuation marks help clarify the intonation, but in written text intonational information is bound to be lost even with punctuation. "I told the doctor I was sick!" would certainly be read with a different intonation if it is incised on a tombstone [Bolinger, 1989, p. 68].

Chafe [1988] has done experiments to explicate the relationship between punctuation and intonation. He claims that there is a "covert prosody" of written language which affects both the writers' and the readers' imagery, and some of this is made explicit by punctuation. His experiments include reading aloud and inserting punctuation to a text from which the original punctuation has been removed. He concludes that *punctuation units* (stretches of language between punctuation marks) can be considerably longer than intonation units.

2.5.3 Text Punctuation

Pascual and Virbel [1996] analyse paragraphing, indentation, and font changes in text understanding and generation, from a semantic point of view. They call certain entities (such as chapters, introductions, theorems) *textual objects* and define a textual architecture by means of meta-sentences that describe the positional, typographical, and speech-act based relations between those objects distinguished by textual punctuation marks. Pascual [1996] gives a fuller model of how such an architecture tailored for scientific and technical documents can be used in formatted text generation. Hovy and Arens [1991] describe how formatting devices such as footnotes, italicized regions, etc. can be planned automatically by recognizing the communicative function of each device. Douglas and Hurst's work characterises layout-oriented devices such as tables and lists [Hurst and Douglas, 1997].

2.6 Summary

There are many dimensions of a linguistic and computational study of punctuation, hinting at a desiderata for a theory of punctuation. The theory should be a unified account of the syntactic, semantic, and discourse effects of punctuation. It should accommodate both structural and text-level punctuation and be formal enough to be applied in the computational analysis and generation of written language. It is hoped that the information-based perspective adopted in this thesis, emphasizing semantic and discourse effects is an estimable try in this sense.

Chapter 3

Linguistic Observations on Informational Effects of Punctuation

3.1 Introduction

In the last chapter, several studies that contain links between semantic and discourse related phenomena (such as discourse relations, intonation) and punctuation have been cited. These studies are generally of speculative nature and do not attempt to characterize the interactions between the phenomena in a unifying manner. For instance, Jones [1997] deals mostly with the syntax of punctuated sentences and rather offhandedly dismisses semantic or discourse effects.

What is being proposed in this thesis is that all these aspects can be seen from an integrated point of view. Structural punctuation in writing contributes to the information structure of a sentence either directly or indirectly by providing cues, cf. Figure 3.1. The non-truth-conditional meaning at sentence level or above is designated by the term *information structure* (to be explained in the next section).

The explanation of information structure and of other semantic or discourse phenomena in its light will clarify the linguistic motivation for modeling punctuated sentences in a computational model. In Section 3.2, brief overviews of the phenomena that are found to be relevant are given. It must be noted that intonation is dealt with in a restricted sense, namely as far as it affects the information structure. The computational aspects of the interaction of intonation and punctuation (as required by text-to-speech generation

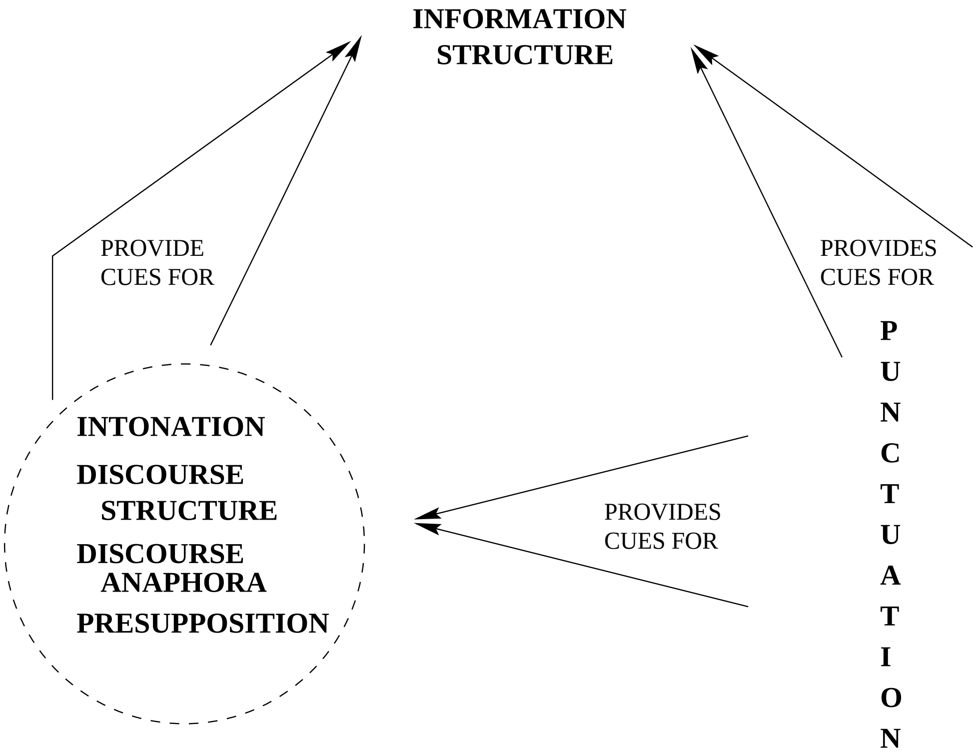

Figure 3.1: Information Structure and Punctuation

systems) are not studied.

Linguistic observations based on corpora are presented in the upcoming sections for dashes, semicolons, colons, and parentheses. These are the four most commonly used structural marks after the period and comma [Jones, 1997, pp. 54–58]. As explained in Chapter 1, comma has already been studied by us [Bayraktar *et al.*, 1998]. Question and exclamation marks' semantic and discourse effects can be said to clearer and better understood than the marks explored in this study.

The motivation for our corpus-based study is to base the formal model of punctuation on English texts from respectable sources. Several factors may influence the information structure of the orthographic sentence: Whether syntactic patterns affect informational status; whether clauses or segments separated by punctuation marks display certain discourse relations; whether anaphoric binding or presuppositional accommodation change by means of punctuation. Punctuated sentences in English from several corpora (WSJ, BC, and BNC) are examined semi-automatically. Computer scripts are written to select relevant sentences and their syntactic patterns.

3.2 Situating Punctuation in Information Structure

Information structure at the sentence level is the non-truth-conditional meaning of a sentence and how it is brought about. Vallduví and Engdahl [1996, p. 460] define the same concept with a different term, *information packaging*, as follows:

> Information packaging is a structuring of sentences by syntactic, prosodic, or morphological means that arises from the need to meet the communicative demands of a particular context or discourse.

The term *packaging* was first used by Wallace Chafe [Engdahl and Vallduví, 1996, quoted in p. 460]:

> ... packaging has ... to do primarily with how the message is sent and only secondarily with the message itself ....

Vallduví [1992, p. 2] gives the following example :

(3.1) a. He hates broccoli.

b. Broccoli he hates.

(3.1a) and (3.1b) are truth-conditionally equivalent but they say what they claim about the world in different ways, the former emphasising an attitude whereas the latter emphasising what is being hated. What is being emphasised corresponds to the new information in a sentence (*focus*) and the rest that links the sentence to the context corresponds to *ground*.

Informational focus of a sentence is the informative (new) part of a sentence that makes a contribution to a reader's mental store. Intonational focus, on the other hand, indicates intonational prominence denoted by any constituent that bears a pitch accent. In English, a subset of the informational focus is realized in situ by intonational prominence [Hendriks, 1996].

At a level higher than the information structure of the sentence is the information structure of a discourse. This comprises the *informativity* and *coherence* of a text, as described in [de Beaugrande and Dressler, 1986, pp. 3–14]. Informativity is the degree to which the occurrences of the text are expected vs. unexpected (or known vs. unknown). Coherence is the pattern in which the units of the text are mutually accessible and relevant.

One way of commenting on informativity and coherence at the discourse level is to specify the discourse relations and their structure. We are inspired by the Rhetorical Structure Theory (RST) [Mann and Thompson, 1987], a proposal about discourse relations between text units. The original study contains characterizations of 25 relations derived from an analysis of 400 texts of varying genres and contexts, by human analysts. Some relations are paratactic (such as Contrast) and span text units of equal importance; some are hypotactic and hold one essential component (nucleus) and a less essential one (satellite). (See Appendix A for RST relations.) An interesting claim of RST is that the same relations that hold between larger segments of the text involved also hold between individual clauses. Text units separated by punctuation marks provide supporting evidence for this claim.

We took a subset of 10 to 12 relations of RST (see [Asher, 1993] and [Corston-Oliver, 1998]) and tested this subset for frequency of occurrence on text units separated by punctuation marks to see if certain marks imply certain relations. Computationally such a study can be used for acquiring heuristic cues (in addition to using discourse markers such as *although, because, as*) for discourse analysis components that do rhetorical parsing [Corston-Oliver, 1998, Marcu, 1997].

A major hypothesis of this thesis is that orthographic means such as punctuation are in several ways contributory to the information structure.

3.2.1 Anaphora

Anaphora is the general mechanism of pointing back within spoken/written discourse either intra- or intersententially to individuals, objects, events, times, and concepts mentioned previously. Anaphora resolution connects an entity to its intended referent by locating a relevant antecedent in the previous discourse. In this thesis, anaphora is taken to be discourse (intra-sentential) anaphora; after all an orthographic sentence (a text-sentence) can include more than one lexical sentence. In (3.2) *He* is a discourse anaphor (a pronominal anaphor) that refers back to *A man*.

(3.2) A man walks. He is wearing a hat.

Pronominal anaphora and its interaction with punctuated text are investigated in the hope that our findings may be used as heuristics. NLP programs already employ, for example, frequency counting and lexical iteration to achieve a recall ratio (the ratio of anaphoric bindings found to those that exist) of 60% in general texts [Mitkov and Boguraev, 1997]. Taking anaphoric cues from punctuation marks may be not only beneficial to improving that ratio but also linguistically interesting on its own right.

3.2.2 Presupposition

The meaning of the term *presupposition* may vary when one moves from semantics to pragmatics [Beaver, 1997, Seuren, 1994]. In semantics, if the truth of one sentence is a condition for another sentence to have a truth value, then the latter is said to presuppose the former. In pragmatics, a speaker's presuppositions are those aspects of an utterance that are taken for granted to be common (mutual) knowledge. Definites ("the King of France"—presupposition: France has a king) and wh-questions ("Which of your sisters is married?"—presupposition: you have more than one sister) are two of the presupposition triggers.

Accommodation is a term coined by Lewis [1979] to refer to the fact that some presuppositions are not uttered explicitly before they are made, but rather reconciled by the hearer post hoc [Seuren, 1994]. For example, if a woman with unknown marital status utters "My husband will be coming in a minute", the hearer accommodates the fact that she is married. It will be shown in the upcoming sections that when a writer is not sure that one of the presuppositions in the written sentence can be accommodated, punctuation marks can act as a device for ensuring that.

3.3 Observations on Dashes

3.3.1 Syntactic Patterns

An obvious question that comes to mind is whether syntactic occurrences of dash usage tend to concentrate on certain patterns [Say and Akman, 1998b]. The second question is whether such patterns relate to the phenomena noted in the previous section. To this effect, dashed sentences were classified according to their patterns using the *whole*[1] of the parsed and tagged version of the WSJ. The results are reflected in Table 3.1.

Syntax Patterns	**No. of Sentences**	
*—NP eos	384	23.64%
—NP— eos	229	14.10%
*—(ADVP|PP|CC) (,) NP(— *) eos	50	3.08%
*—S eos	149	9.17%
—SBAR— eos	74	4.56%
—S— eos	69	4.25%
—PP— eos	121	7.45%
*—PP eos	77	4.74%
VP-related	95	5.85%
Other	376	23.15%
TOTAL	**1624**	**100%**

Table 3.1: Distribution of Syntax Patterns for Dashes (*Other* row includes various low frequency patterns)

Except for the WSJ-specific usages, most of the dash usage (about 70%) relate to the use of noun phrases (NP-related) and lexical sentences or sentential complements (S-related). More specifically, mid-sentence or end-of-sentence noun phrases, mid-sentence prepositional phrases, and sentences or sentential complements that come at the end are most common. Not only that the distribution patterns of dashed sentences are quite stable but also, as will be seen in the next subsection, both common patterns and some less common ones relate to the semantic and discourse phenomena in interesting

[1]Syntactic classifications (Table 3.1) have been done on the complete WSJ while discourse structure related ones (Table 3.2) span a mixed subset of the indicated corpora.

ways. For example, a NP that comes at the end of a sentence (first row of Table 3.1) is usually informationally prominent. A sentence or a sentential complement in the middle of the orthographic sentence (sixth and seventh rows) usually indicates a Commentary relation or Apposition whereas a similar constituent conjoined by a single dash (fourth row) signals for relations such as Elaboration, Contrast, or Parallel. The syntax of the patterns can be taken as a guide in a computational system (see Section 2.4), but it will be explored in this thesis as far as it acts as a pointer for semantic and discourse phenomena.

3.3.2 Constraints on Discourse and Information Structure

Dashes cue for a certain distribution of discourse relations. In Table 3.2, the discourse relations that are found between units separated by dashes are shown. Note that although there is a distribution pattern, these relations are due to the semantics and context of the sentences. Thus their distribution could be slightly different for another interpreter who tries to classify them into the same set of relations. Since our aim is to make overall characterizations rather than calculate strict distributions with respect to discourse relations, such a difference is immaterial. The relations used by Asher [Asher, 1993, pp. 302–304] are taken as a basis (see Appendix A for details).

The following observations are due to examining around 400 randomly selected sentences with dashes, a subset of which (namely, 125 sentences) were studied for discourse relations only. The sentences come from the WSJ from Penn Treebank 2 [Marcus *et al.*, 1993], the BNC [BNC, 1997], and BC [Kučera, 1992] (and a subset of it called SUSANNE [Sampson, 1995]).

The row denoting other usages in Table 3.2 consists of corpus-specific reference mechanisms, title introduction, list introduction using dashes, or one-off usages such as introducing quoted sentences. As can be seen from the table, the distribution of discourse relations in sentences with dashes is not completely ad hoc and is worthy of special consideration. Indeed, 56% of the relations are in the categories of Elaboration, Commentary, and Apposition.[2]

[2]The reason for treating Apposition as a kind of discourse relation when used in conjunction with dashes is that it is usually a special case of Apposition with emphasis. This effect is more clearly seen

Discourse Relations	**No. of Sentences**	
Elaboration	26	20.8%
Commentary	20	16%
Apposition	24	19.2%
Explanation	12	9.6%
Contrast	6	4.8%
Parallel	3	2.4%
Result	2	1.6%
Instance	3	2.4%
Continuation	2	1.6%
Cause	2	1.6%
Background	1	0.8%
Other	24	19.2%
Total	125	100%

Table 3.2: Distribution of Discourse Relations for Dashes

The second observation relates to an interesting function of dashes: they can be used to denote *focus*—in a combination of the informational [Vallduví, 1993] and the intonational senses. Some dashes do not disrupt the syntactic flow of the sentence; in other words, they solely add an element of emphasis. This could indicate an extra level of emphasis on informational prominence, where an intonational focus would already be expected in spoken text (see (3.3)), or distinguish what would have been an intonational focus on a lexical word or phrase in speech (see (3.5), (3.6), (3.4)). These patterns correspond respectively to a coordinating conjunction (CC) such as *and* followed by a noun phrase (NP); an adjectival phrase (ADJP); or an adverbial phrase (ADVP).

(3.3) **(WSJ)** Already, the consequences are being felt by other players in the financial markets—even governments.

(3.4) **(BC)** To understand the past history—and the future potential—of American Catholic higher education, it is necessary to appreciate the special character of

if one thinks of the possibility of substituting another mark (or marks) in place of the dash(es). In the case of Apposition, dashes can be replaced by commas while incurring some loss of emphasis in most of the occurrences.

the esprit de corps of the religious community.

(3.5) **(WSJ)** Knowing a tasty—and free—meal when they eat one, the executives gave the chefs a standing ovation.

(3.6) **(BC)** Fifteen members of the Republican State Committee who are retiring—voluntarily—this year were honored yesterday by their colleagues.

When the dash comes at the end of the sentence it is usually more prominent informationwise than its mid-sentence counterparts. This might be due to the fact that it is cognitively more plausible for the human mind to consume the information acquired most recently [Sperber and Wilson, 1986]. Compare (3.7) with (3.8) as examples of changing prominence. The pattern that comes at the end of the sentence usually corresponds to an end-of-sentence noun phrase or a sentential complement.

(3.7) **(WSJ)** In addition, the Cray-3 will contain 16 processors—twice as many as the largest current supercomputer.

(3.8) **(WSJ)** Some of the biggest service-industry exporters—American financial-service companies, for example—have yet to be fully included in our export statistics.

Some styles of writing (e.g., a particular brochure by a health organization, as found in the BNC) make repeated use of the above effect and employ dash interpolations for intonational focus to keep a vivid and striking pace throughout the document.

An end-of-sentence dash might convey key information in that the following unit gives out some information otherwise not mentioned overtly in the sentence. In such cases, intonational focus and part of informational focus fall on the dash interpolation (see (3.9), (3.10)). This kind of usage can syntactically correspond to apposition or what has been termed as *right dislocation* [Hadumod, 1996, p. 134], viz. the appearance of syntactic constituents at or outside the right boundary of the sentence where the original position is marked by a pronominal element (see (3.9)). This usage syntactically corresponds to a noun phrase at the end of a sentence.

(3.9) **(BC)** This is largely because of the unpredictability of the man who operates the helm of the state government and is the elected leader of its two million inhabitants—Gov. Ross Barnett.

(3.10) **(WSJ)** As a result, marketers of faux gems steadily lost space in department stores to more fashionable rivals—cosmetics makers.

Even when they are part of informational focus, mid-sentence dash-enclosed units can be parenthetical and less prominent than other parts of the sentence. They can provide background or extra information, and comments that are not necessarily crucial to the understanding of the sentence (see (3.11), (3.12)). The syntactic pattern usually corresponds to a sentence.

(3.11) **(WSJ)** The department said orders for nondurable goods—those intended to last fewer than three years—fell 0.3% in September to $109.73 billion after climbing 0.9% the month before.

(3.12) **(WSJ)** Still, the restaurant's ever-changing menu of five-course dinners—it supposedly hasn't repeated a meal since opening in 1971—requires constant improvisation.

On the other hand, dash-enclosed units can also change the perspective of the reader by offering an alternative wording, e.g., (3.13), (3.14), (3.15). Within the dashes, the reader is directed to a different encyclopedic entry in a relevance-theoretic way. That is, the writer uses the dash interpolation as a means to establish the maximum contextual effect with minimal processing effort for the reader by overriding or strengthening the meaning of the lexical entry it is adjoined to [Blakemore, 1996, Sperber and Wilson, 1986]. This effect arises with verb phrase (VP) related constituents.

(3.13) **(WSJ)** They showed up, but didn't—or couldn't—challenge.

(3.14) **(WSJ)** Ogilvy under the fastidious Mr. Roman gained a reputation as occasionally being high-handed in its treatment of clients, of preaching what strategy a client should—indeed, must—follow.

(3.15) **(WSJ)** "I agree, it's ridiculous," says Mr. Boren, and indeed by now ridiculous may be the only way to describe how the U.S. decides to take—or rather, not to take—covert action.

3.3.3 Constraints on Anaphora and Presupposition

The next question to consider is whether these observations at the discourse level have implications for anaphora resolution. The basic observation is that antecedents within dashes are less felicitous if the part enclosed in dashes has an adjoining (parenthetical) status and is mid-sentence (except when the antecedents introduced within the dashes form an apposition to the noun phrase that they are adjoined to). This is not so with conjoining status dashes where other factors (grammatical function, lexical iteration, etc.) function as normal.[3]

In (3.16) this observation seems to have been violated. Native speakers found *these countries* to be ambiguous as to which countries it included.

(3.16) **(WSJ)** "If America can keep up the present situation—her markets open for another 15 years, with adjustments, and Japan can grow and not cut back, and so too, Korea, Taiwan, Hong Kong, Singapore, ASEAN,[4] Australia and New Zealand—then in 15 years, the economies of these countries would be totally restructured to be able to almost sustain growth by themselves."

On the other hand, in (3.17), *their parents* does not stand as a felicitous candidate for further anaphoric reference, though it stands in the subject position (a strong position to be an anaphoric candidate) from within the part enclosed in the dashes.

(3.17) **(WSJ)** The issue is further complicated because although the organizations represent Korean residents, those residents were largely born and raised in Japan and many speak only Japanese. That they retain Korean citizenship and ties is a reflection of history—their parents were shipped in as laborers

[3]In the former case, where there is a parenthetical dash interpolation, other factors of anaphoric reference as depicted by the centering framework [Grosz *et al.*, 1995, Turan, 1997] still continue to function. The existence of the parenthetical may serve as a preference or overriding factor.

[4]ASEAN stands for the Association of South East Asian Nations.

during the decades when Japan occupied Korea before World War II—and the discrimination that still faces Koreans in Japanese society.

Some intuitive yet made-up sentences will be used to make this point clearer. The first sentence of each of the following examples is taken from corpora. The second sentence in each case is invented by us while an attempt is made preserve coherency.[5]

In (3.18), the pronominal anaphor in the second sentence (*They*) can be resolved easily, whereas in (3.19), the resolution of *They* to *outside observers* seems to be non-felicitous.

(3.18) **(WSJ)** The hazardous waste is growing on Mr. Courter's property—the neighbors are suing for consumer fraud. They are ready to fight till the end.

(3.19) **(BNC)** We know from experience that many factors—some of which may never be apparent to outside observers—determine whether a prisoner of conscience is released. They may even consider the wrong factors.

A complementary hypothesis worth looking at is thus as follows: When a dash precedes the rightmost constituent and falls on a lexical phrase (usually an NP), any discourse referent introduced in that phrase (see (3.20) where the second sentence is again made-up by us) is a more salient choice than it would otherwise be for serving as an antecedent in the next sentence. The resolution of *They* to *cosmetic makers* seems to be a more felicitous choice unless the context dictates otherwise.

(3.20) (=3.10) **(WSJ)** As a result, marketers of faux gems steadily lost space in department stores to more fashionable rivals—cosmetics makers. They are really aggressive.

Use of mid-sentence dashes for accommodating presuppositions, on the other hand, often comes in the form of appositions, see (3.21), (3.22), (3.23). Some dash-interpolations are used to clarify the presuppositions of the constituent they are adjoined

[5]The weak existence of such pairs in corpora may be due to the fact that the work was conducted in a rather small corpus of dashed sentences and that dashes are not used in such a widespread way. Dashes constitute normally 2-5% of punctuation marks [Jones, 1997].

to when it is not clear to the writer that the constituent itself is in the domain of shared knowledge with the reader.

(3.21) **(WSJ)** That can be a trap for unwary investors, says Richard Bernstein, senior quantitative analyst at Merrill Lynch & Co. Strong dividend growth, he says, is "the black widow of valuation"—a reference to the female spiders that attract males and then kill them after mating.

(3.22) **(WSJ)** Both the British Diabetic Association and the Committee on Safety in Medicines—Britain's equivalent of the U.S. FDA—recently issued statements noting the lack of hard scientific evidence to support Dr. Toseland's findings.

(3.23) (=1.2) **(BC)** It is a killer sub—that is, a hunter of enemy subs.

3.4 Observations on Semicolons

Syntactic patterns for semicolons are considerably less varied than the patterns for dashes, cf. Table 3.3 which reports data for the whole of WSJ. The most common patterns involve joined sentences (first row) or noun phrases (second row). There is also a non-negligible amount of usage having a conjunction before the last constituent joined by a semicolon. The first pattern may cue for a variety of discourse relations between the units separated by semicolons. The second pattern occurs usually when the semicolons act as a syntactic separator. The third and fourth rows of Table 3.3 might denote special informational focus or a topic change as will be seen in examples.

Table 3.4 depicts the usage of semicolons in relation with discourse sentences studied separately for the WSJ and the BNC, with 200 randomly selected sentences from each. In the WSJ, it is more common to regard semicolons as syntactic separators. This difference may be due to the financial content of many of the sentences in the WSJ. On the other hand, the sentences chosen from the BNC—coming from brochures and art and literary criticism—displayed a tendency to mark a continuation relation between sentences or verb phrases. Elaboration relation occurs between text units separated by semicolons, with frequent uses of Contrast, Parallel, and Instance following.

Syntax Patterns	**No. of Sentences**	
S(;S)+ eos	548	58.17%
NP(;NP)+ *	243	25.80%
S(;S)+;CC S eos	40	4.25%
NP(;NP)+; CC NP *	64	6.80%
VP(;VP)+ *	10	1.06%
PP(;PP)+ *	4	0.42%
Others	33	3.50%
TOTAL	**942**	**100%**

Table 3.3: Distribution of Syntax Patterns for Semicolons

Pronominal anaphora in the subject position of a second clause of a semi-coloned sentence resolve to the subject position of the first clause because of the increased coherence. This is already a well-known heuristics in anaphora resolution [Kennedy and Boguraev, 1996, Mitkov and Boguraev, 1997]; but it should have a stronger weight in case of sentences joined with semicolons (see (3.24), (3.25), (3.26) for italicised pairs corresponding to anaphora resolution).

(3.24) **(WSJ)** *Examiners from the Office of the Comptroller* of the Currency had been combing through First Interstate's real-estate portfolio since last month; *they* first recommended that First Interstate take a provision that was less than the eventual $350 million third-quarter hit. [Emphases added.]

(3.25) **(WSJ)** *Itel* bought a 17% stake in Sante Fe Pacific last year and *Olympia & York* later purchased about a 20% stake; *they* would have interests in the new realty company in line with their holdings in Sante Fe Pacific. [Emphases added.]

(3.26) **(BNC)** *English fiction* loves such people; *it* never tires of the lurch, of such areas of darkness. [Emphases added.]

Patterns such as the third pattern of Table 3.3 imply that the informationally prominent part of the sentence is after the connective (e.g, *and*). This might imply a change of

Usage	**No. of Sentences**			
	WSJ	**BNC**	**Total**	**Percentage**
Continuation	22	61	83	20.75%
Elaboration	25	21	46	11.50%
Contrast	16	21	37	9.25%
Parallel	18	17	35	8.75%
Instance	9	18	27	6.75%
Commentary	5	13	18	4.50%
Explanation	6	12	18	4.50%
Result	8	7	15	3.75%
Cause	5	4	9	2.25%
Introduction	1	5	6	1.50%
Background	3	-	3	0.75%
Purpose	1	-	1	0.25%
Other	81	21	102	25.50%
Total	**200**	**200**	**400**	**100%**

Table 3.4: Distribution of Discourse Relations for Semicolons (*Other* row includes syntactic separators.)

topic as in (3.27) and (3.28) or informational prominence as in (3.29). The punctuational structure can be emphasised by a discourse marker (*naturally*) as in (3.30).

(3.27) **(WSJ)** The weight of Lebanon's history was also against him; and it is a history Israel is in danger of repeating.

(3.28) **(BNC)** Occasionally a book has almost achieved immortality, like John Ruskin's Stones of Venice, but even more modest books can call up the spirit of a place; and a private letter may illuminate both a person and a work of art. [*Discourse continues with the topic of private letters.*]

(3.29) **(WSJ)** There are many reasons for the market's jumpiness: new trading vehicles such as stock-index futures and options; computer-driven strategies like program trading; and crowd psychology.

(3.30) **(WSJ)** The club plans to show nerdy movies, such as "Real Genius," in which physics whizzes pop corn with lasers; and naturally, the "Revenge of the Nerds," a tale of college males with runny noses and ill-fitting pants.

With certain discourse relations (such as Parallel, Contrast, Continuation) semicolons may cue a special temporal or spatial proximity (see (3.31) for parallel parts; (3.32), (3.33) for close temporal continuity), or at least a special relation between spatiotemporal aspects if not proximity (see (3.34) for contrasting events related with passage of time).

(3.31) **(WSJ)** Anti-Jones sentiment flooded the local press: "A crude obnoxious hick," said one writer; "a real oink," said another; "Who in the hell does he think he is?" wrote a third.

(3.32) **(WSJ)** By noon, Mr. Bush had taken two phone calls from Vice President Dan Quayle, who was in California; made a televised statement of concern; signed a disaster proclamation; received a written report from the Federal Emergency Management Agency; and visited FEMA headquarters.

(3.33) **(WSJ)** Prosecutors, in an indictment based on the grand jury's report, maintain that at various times since 1975, he owned a secret and illegal interest in a beer distributorship; plotted hidden ownership interests in real estate that presented an alleged conflict of interest; set up a dummy corporation to buy a car and obtain insurance for his former girlfriend (now his second wife); and maintained 54 accounts in six banks in Cambria County.

(3.34) **(BNC)** With the rise of modernism, Rodin's reputation fell; with the decline of modernism, Rodin's fame is growing again.

Pairs of semicoloned sentences or verb phrases have closer coherence than those surrounding them (see (3.35), where semicoloned sentences list a closely related set of parental activities, and (3.36), where the series of events that took place are closely related with an accummulated effect).

(3.35) **(WSJ)** Parents should be involved with their children's education at home, not in school. *They should see to it that their kids don't play truant; they should* *make certain that the children spend enough time doing homework; they should scrutinize the report card.* If parents are dissatisfied with a school, they should have the option of switching to another. [Emphases added]

(3.36) **(WSJ)** But there are times when they seize up, and panicky sellers cannot find buyers. That's just what happened in the October 1987 crash. *As the market tumbled, disorderly market conditions prevailed: The margins between buying bids and selling bids widened; trading in many stocks was suspended; orders took unduly long to be executed; and many specialists stopped trading altogether.* These failures in turn contributed to the fall in the market averages. [Emphases added]

Also worth mentioning are special effects that are borne out of the combination of two or more marks, e.g., the dialogue-like nature of (3.37).

(3.37) **(BC)** Inspiring—yes; instructive—maybe; duplicable—no!

Such uses are powerful in emphasizing meaning but are not common.

3.5 Observations on Colons

In over 70% of the sentences of the WSJ including a colon (see Table 3.5), either an NP or a sentence-related constituent follows the colon. The most common category is an NP followed by another NP separated by a colon. This again might stem from the specific nature of the WSJ as it has lots of sentences presenting financial data. It will be observed in the following examples that an NP coming just before an end-of-sentence marker might denote informational and intonational focus. A sentence after a colon can be related to the preamble to the colon with a variety of discourse relations.

Table 3.6 depicts the usage of colons within a subset of WSJ and BNC. As can be seen, the difference in usage between two corpora is not significant for the sample set of 400 sentences. Only about 15% of the sentences involve relations other than Introduction. Yet again as in the case of Apposition, Introductions can cooccur with discourse relations such as Elaboration and Instance. *Other* uses include speech and list introduction, representing titles and time.

Syntax Patterns	**No. of Sentences**	
S:S eos	237	14.24%
NP:S eos	219	13.16%
VP:S eos	165	9.92%
NP:NP eos	517	31.07%
S:NP eos	24	1.44%
Lists	157	9.44%
WSJ Headlines	156	9.38%
*:PP eos	82	4.93%
Other	107	6.43%
TOTAL	**1664**	**100%**

Table 3.5: Distribution of Syntax Patterns for Colons

Colons, by a great majority, are used as a means of introduction (e.g., to introduce concepts, lists, speech), see (3.38). They can also be used to introduce one concept, object, etc. from a set of similar ones, see (3.39).

(3.38) **(WSJ)** One sign of Mr. Deaver's renaissance: an appearance on ABC's "Nightline" for a show on pack journalism.

(3.39) **(BNC)** A good defence lawyer would now be armed with all the mitigating circumstances of your life: mental records, character witnesses, ... any reason why your life should be spared.

Alternatively, the preamble of the sentence until the colon can build up some expectancy on the reader's part towards introducing the idea, object, etc. mentioned in the follow-up:

(3.40) **(WSJ)** Some of our best and most idiosyncratic film makers—from Truffaut to Fellini to Woody Allen—have taken a cue from Chekhov: When it comes to compelling drama, there's no place like home.

(3.41) **(WSJ)** Many small investors are facing a double whammy this year: They got hurt by investing in the highly risky junk bond market, and the pain is worse because they did it with borrowed money.

Usage	**No. of Sentences**			
	WSJ	**BNC**	**Total**	**Percentage**
Introduction	122	93	215	53.75%
Instance	4	7	11	2.75%
Elaboration	5	5	10	2.50%
Explanation	4	3	7	1.75%
Representing Time	5	2	7	1.75%
Cause	4	-	4	1.00%
Continuation	-	4	4	1.00%
Result	1	2	3	0.75%
Contrast	1	1	2	0.50%
Parallel	-	1	1	0.25%
Other	59	84	143	35.75%
Total	**200**	**200**	**400**	**100%**

Table 3.6: Distribution of Discourse Relations for Colons

(3.42) **(WSJ)** But what happens next in the continuing takeover drama may depend more on the company's two most powerful and fractious unions: the pilots and machinists.

Such usages, especially when they introduce an NP, contain part of the informational focus of the sentence. An NP introduced in such a way can be given preference in selecting an antecedent for a following pronoun. The discussion for a similar usage in dashes applies here as well. Thus, *They* would preferably be resolved to Soviets who travel abroad in (3.43).

(3.43) **(WSJ)** Unless other rules are changed, the [officials] could cause difficulties for the people [the devaluation] is primarily meant to help: Soviets who travel abroad. *They would look for ways of earning hard cash.* [*The second sentence added*][6]

However, this observation is again context-dependent. In (3.44) the individual nouns

[6]The original sentence ("Unless other rules are changed, the devaluation could cause difficulties for the people it is primarily meant to help: Soviets who travel abroad.") has antecedents easier to distinguish as *devaluation* is inanimate.

introduced in the NP after the colon form the parts of a concept in the preamble to the colon and are unlikely to be referred to separately.

(3.44) **(BNC)** In a book called How to Appreciate Pictures by R. C. Witt, written in 1902, the chapter headings are not so different: drawing, colour, light and shade, composition, treatment, methods and materials.

When an NP introduces an NP, the style is emphatic and compact, and harder to detect computationally as it consists of fragments rather than full sentences. Heuristics could play a role here again; for example, in sentence (3.46), when a place name is detected before a colon, it could be noted (by means of a special lexicon) as the place where the following event took place.

(3.45) **(WSJ)** Employees were told that if they really wanted the publications, they would have to have them sent home instead. The reason: overload, especially of non-subscription magazines.

(3.46) **(BNC)** Santiago, Chile: On a warm autumn evening in 1990 international rock star Sting dances on stage with a group of Chilean mothers and grandmothers of the "disappeared". [...] Moscow, USSR: At the same time, 12,000 miles away, a delegation from Amnesty's International Secretariat is making plans for a fledgling Moscow Group to participate in the Women in the Front Line campaign.

Such heuristics could be useful in certain computational applications. Text summarization is one where text-adjuncts introduced by columns could be extracted as key concepts. An anaphora resolution component is another where such constituents could be given preference as canditate antecedents.

Other discourse usages include elaboration of a concept introduced in the preamble to the colon, see (3.47), (3.48).

(3.47) **(WSJ)** A volcano will erupt next month on the fabled Strip: a 60-foot mountain spewing smoke and flame every five minutes.

(3.48) **(BNC)** The Commission also took a stronger stand in respect of two other countries on its agenda: Cuba will not come under special scrutiny by a representative of the UN Secretary-General and the Expert on Equatorial Guinea, a country which receives assistance under the UN Advisory Services Program, has been requested to study the human rights situation there.

Other sentential patterns might involve an instance (exemplification) of an eventuality, see (3.49), (3.50).

(3.49) **(WSJ)** And he isn't the only big spender: NBC will pay a record $401 million for the 1992 Summer Games, and ESPN, 80%-owned by Capital Cities/ABC, will shell out $400 million for four years of baseball, airing 175 regular-season games a year.

(3.50) **(BNC)** Deaths in detention are not phenomena of the distant past: last year two people were reported as having died in custody.

Alternatively, the part following the clause might attribute a reason or explanation for the preamble, see (3.51), (3.52).

(3.51) **(WSJ)** In part, this division is dictated by economics: West Germany is a net exporter of capital while the U.S. isn't.

(3.52) **(BNC)** The jury will be "death qualified" (in the case of Alabama capital law): all jurors have to believe in capital punishment, and be prepared to sentence someone to death.

In some of the sentences, where the sequel to the colon is a quotation, there may be a change of reference systems (for example, from third person to first person), as exemplified by (3.53), (3.54). Quotation marks or textual devices such as indentation of paragraphs make up a more reliable signal for such a change. However, the latter kind of devices might be lost to inadequate annotation of the corpus. Even for the case of known quoted speech, Kennedy and Boguraev [1996] report problems in trying to resolve anaphora.

Syntax Patterns	**No. of Sentences**	
* (NP) *	904	63.48%
* (S (eos S)+) *	194	13.62%
* (ADJP) *	46	3.23%
* (PP) *	44	3.09%
* (CC *)	36	2.53%
* (VP) *	32	2.25%
* ("S"|"NP") *	23	1.62%
* (ADVP) *	8	0.56%
Other	137	9.62%
TOTAL	**1424**	**100%**

Table 3.7: Distribution of Syntax Patterns for Parentheses

(3.53) **(BNC)** Leslie was himself a painter, and thus comments in his own right on his friend's art, as here on a painting of Hampstead Heath: "I have before noticed that what are commonly called warm colours are not necessary to produce the impression of warmth in landscape [...].

(3.54) **(BNC)** Lawrence, for example, was in Parma in 1820, in ecstasy over the work of Correggio: "Four times I went on long visits to the St Jerome, his finest work".

3.6 Observations on Parentheses

As can be seen from Table 3.7, parentheses mostly enclose noun phrases, with sentences or sentential complements making the other possible candidates. There is also a subtantial number of miscellaneous uses with different syntax patterns occurring within the parentheses. The NPs enclosed can be directives or missing words as will be seen in the examples and are always parenthetical in sense (except when preceded by coordinating conjunction (e.g., *and*)). There can also be more than one sentence enclosed within parentheses.

Table 3.8 depicts the usage of parentheses within a subset of the WSJ and the BNC.

Appositives and non-discourse related uses such as directives (references to sources), missing words, etc. are the most frequent uses. This is not surprising since mostly noun phrases are enclosed within the parentheses. Still, some of the uses exhibit a variety of relations with respect to the sentence the parenthetical follows.

Usage	**No. of Sentences**			
	WSJ	**BNC**	**Total**	**Percentage**
Apposition	87	45	132	33.00%
Commentary	11	28	39	9.75%
Elaboration	20	13	33	8.25%
Background	9	1	10	2.50%
Contrast	-	9	9	2.25%
Instance	-	8	8	2.00%
Intonation	2	5	7	1.75%
Cause	2	3	5	1.25%
Continuation	3	1	4	1.00%
Introduction	-	2	2	0.50%
Condition	-	1	1	0.25%
Parallel	1	-	1	0.25%
Other	65	84	149	37.25%
Total	**200**	**200**	**400**	**100%**

Table 3.8: Distribution of Discourse Relations for Parentheses

Parentheses, even when they occur at the end of a sentence, are used in a parenthetical sense and come in pairs. Dashes, although they usually carry a similar parenthetical sense when used as adjuncts, denote more prominent entities when they come at the end of the sentence and are used in a conjunctive sense.

In (3.55), a pronoun referring to Lebanon instead of Ethiopia is rendered unlikely to occur later in the text.

(3.55) **(WSJ)** Recently, in Wollo province in the center of Ethiopia, Tigrean forces have killed, wounded and captured an additional 20,000 government troops. (Think what these numbers mean—considering the headline space devoted to hundreds of deaths in Lebanon, a small country of little strategic importance!)

Also, within a pair of parentheses there may be more than one sentence. This does not happen with other "parenthetical" marks (i.e., dashes and commas). In the case of multiple sentences within parentheses, that segment of the discourse will act as a parenthetical at discourse level; this is also known as a *discourse pop* [Asher, 1993, p. 279]. Normal restrictions for accessibility of anaphors apply within the enclosed segment. However, for the text following the parenthesized constructs the referents introduced within are defeasibly inaccessible. This condition of defeasible inaccessibility is stronger for parentheses than for dashes. In (3.56), the sentences enclosed in parentheses give background information about a certain company mentioning its previous name. However, the parentheses make the old name parenthetical in nature and defeasibly inaccessible in the rest of the discourse. There could even be an embedded level of defeasibility as can be seen in (3.57) where dashes create a further level of parentheticals within parentheses.

(3.56) **(WSJ)** Mr. Stein said he expects profit to be higher in 1990 than in 1989, reflecting a number of measures taken since the acquisition of Ekco Housewares in late 1987. (Prior to acquiring the housewares business, the company was known as Centronics Corp.; Centronics had been a maker of computer printers, but Mr. Stein and other officers decided to sell that business after Japanese competitors grabbed a dominant share of the market.)

Next Sentence: Mr. Stein said tighter operating controls have enabled Ekco to reduce inventory levels 25% to 30%; improve on-time delivery of orders to about 95% from around 70%; and to lower the number of labor hours required to produce a unit.

(3.57) **(WSJ)** Mr. Schwartz, the puckish planner from Englewood, Colo., says that allowing the business to police itself would be "like putting Dracula in charge of the blood bank." Mr. Gargan, the Tampa planner who heads one trade group, favors simply assessing the industry and giving the money to the SEC to hire more staff.

(Mr. Gargan's views are not greeted with wild enthusiasm over at the IAFP, the major industry organization. [...] Then he sent the pooch's picture with the certificate of membership—it was made out to "Boris 'Bo' Regaard"—to every newspaper he could think of.)

Next Sentence: The states have their own ideas about regulation and certification.

Somewhat like the dashes, parentheses may enclose intonationally prominent items as well as appositions, commentaries, background items, and several other discourse-related items. As distinct from the dashes, they also enclose missing words or phrases (see (3.58), (3.59)) and directives (a general term which is used here to denote abbreviations, references, and sentences directing the attention of the reader to another item, as in (3.60)).

(3.58) **(WSJ)** "We've done a lot to improve (U.S.) results and a lot more will be done," Mr. Mark said.

(3.59) **(WSJ)** One example he gives: "She didn't ask" (why the Palestinian children are soldiers throwing stones).

(3.60) **(WSJ)** Mr. Titus is a researcher at the Justice Department's National Institute of Justice. (See related story: "Small Merchants' Big Burdens" – WSJ Oct. 23, 1989)

Parentheses are also used pre-appositionally more often than dashes, e.g., to denote an apparently obvious connotation of a word to make sure it is understood right (see (3.61), (3.62) and the first sentence of (3.63)). They are also used to denote presuppositions that must be explicitly stated for a majority of the readers, see (3.63), the second sentence.

(3.61) **(BNC)** In the same way, if your spouse pays income tax at the higher rate and you pay tax at only the (lower) basic rate, then in order to obtain Higher Rate Tax Relief, your spouse should enter into the covenant, or into a Joint Deed of Covenant with you, [...].

(3.62) **(WSJ)** "Most of our competitors are announcing products based on our (older) products," said Finis Conner, chief executive officer and founder of the firm that bears his name.

(3.63) **(WSJ)** To wit, my maiden voyage (and novitiates are referred to as virgins) began at dawn on a dew-sodden fairway and ended at noon in a soggy field. (Balloon flights almost always occur at dawn or dusk, when the winds are lightest.)

In (3.64), the constituent between the parentheses is parenthetical as opposed to being intonationally prominent. In (3.65), the constituent may be deemed to be intonationally prominent. So a conjunction followed by syntactic category such as an NP does not by itself denote intonational prominence.

(3.64) **(BNC)** Perhaps graduates of a number of drama schools might be given a provisional Equity card requiring a minimum number of engagements (and/or weeks) to be worked within the two or three years of its validity [...].

(3.65) **(BNC)** He is unlikely to have lost his distrust of the self; and he is likely (and welcome) to resume his furious fictions.

3.7 Summary

This chapter has outlined (with examples drawn from well-known corpora) a variety of phenomena observed in conjunction with our four punctuation marks: dash, colon, semicolon, parentheses. Data on syntactic patterns of the marks and their relation to discourse relations have also been presented. In the next chapter, the same phenomena will be formally modeled with SDRT.

Chapter 4

An Informational Model for Punctuation

4.1 Introduction

Discourse Representation Theory (DRT) [Kamp and Reyle, 1993], an influential theory dealing with discourse related phenomena, is our framework. In particular, Asher's [Asher, 1993] extension, Segmented Discourse Representation Theory (SDRT), and his application of this theory to abstract entity anaphora[1] proves valuable for our study. Cues coming from the punctuation marks affect the properties of the discourse structure ascribed to the text segment, thus contributing to anaphora resolution, topic identification, etc. [Say and Akman, 1996, Say, 1997, Say and Akman, 1998a]. SDRT has established a standard for a theory of discourse structure where in order to be coherent, constituents of the discourse must be attached together via discourse relations, as inspired by Rhetorical Structure Theory [Mann and Thompson, 1987]. (S)DRSs (Segmented Discourse Representation Structures) are seen by some researchers [Cormack, 1992] as mental models of discourse comprehension. While such a view is promising for further work, SDRT at least provides the starting points for modeling of the role of punctuation marks in text understanding.

[1] "John believes everything Mary says. Further it is all true." [Asher, 1993, p. 48]. This is an example of abstract entity anaphora where "everything Mary says" is referred to by a single entity, *it*.

4.2 An SDRT Model for Punctuation

The definitions of text-sentence, text-clause, and text-phrase are borrowed from Nunberg [Nunberg, 1990]. It will be apt here to digress and offer a slightly revised version of Nunberg's terminology employed in his text-grammar. A paragraph consists of one or more text-sentences (ending with a period, exclamation, or question mark). A text-sentence consists of one or more text-clauses separated by semicolons or two clauses separated by a single conjunctive dash. A text-clause can further contain clausal adjuncts (e.g., colon expansion, dash interpolations) that can come anywhere in the sentence except at the beginning or in immediate adjacency of another clausal adjunct.[2] With the restriction that they are not embedded in another clausal adjunct of same kind, clausal adjuncts can contain one or more text-clauses. A text-phrase is the name for a text-grammar constituent that does not contain a colon expansion. Its correspondent in the lexical grammar can be a lexical sentence, a clause, a constituent such as a noun phrase, or even a fragment.

The definitions and examples of discourse relations used in modeling cues by punctuation will be found in Appendix A. For an overview of DRT and SDRT, the reader is referred to Appendix B. All temporal information has been removed from the (S)DRSs in this chapter so as not to create unnecessary visual clutter.

The proposed interpretations of punctuational cues already described in Chapter 3 can be roughly divided into three.

4.2.1 Information and Discourse Structure

Discourse referents such as those in (4.1) are more prominent than others as they denote special intonational focus. Syntactically, a text-phrase (usually an NP) comes after the punctuation mark under consideration.

(4.1) **(BC)** [simplified] This is the underlying concern along with the lack of time—the shortage of cash. It is an acute problem.

[2] A colon expansion is not arbitrary as to its placement; it must always come at the end.

A modification to SDRT is to denote such an intonational focus with an underlined discourse referent (Figure 4.1).

$$\boxed{\begin{array}{l} \quad\quad \text{x } \underline{\text{y}} \text{ z} \\ \text{lack-of-time(x)} \\ \text{shortage-of-cash(y)} \\ \text{underlying-concern(x)} \\ \text{underlying-concern(y)} \\ \text{z} = \text{y} \\ \text{acute-problem(z)} \end{array}}$$

Figure 4.1: DRS for (4.1)

SDRT could be used to insert such an effect in to the discourse structure. A suitable choice here is topic-based updating, by taking the constituent (here the NP) that denotes special focus and destructively adding it as a summary that dominates the discourse (Figure 4.2).

$$\boxed{\begin{array}{l} k_0{:=}\boxed{\begin{array}{c} \text{y} \\ \text{shortage-of-cash(y)} \end{array}} \\ \quad\quad\quad \Downarrow \\ k_1{:=}\boxed{\begin{array}{l} \quad\quad \text{x z} \\ \text{lack-of-time(x)} \\ \text{underlying-concern(x)} \\ \text{underlying-concern(y)} \\ \text{z} = \text{y} \\ \text{acute-problem(z)} \end{array}} \\ \text{Continuation}(k_0,k_1) \end{array}}$$

Figure 4.2: Revised SDRS for (4.1)

A similar effect is achieved with a colon. In (4.2) (taken from [Quirk *et al.*, 1972, p. 1068]), the NP serves as the summary or topic constituent of the whole SDRS attributed to the sentence.

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

k_0:=	x country-cottage(x)
	$\Downarrow$
k_1:=	u e u = ? e: remaining-desire-above-all-else(u,x)
Elaboration(k_0, k_1)	

Figure 4.3: SDRS for (4.2)

The second modification to SDRT takes into account the alternative wording effect of dash interpolations such as (4.3) ([Blakemore, 1996, p. 116]).

(4.3) They ran—sprinted—up the hill.

The $\triangleright$ sign in the conditions of Figure 4.4[3] shows the overriding effect. Another way

U v e
U = ?
e:run(U) $\triangleright$ e:sprint(U)
the hill(v)
up(U,v)

Figure 4.4: DRS for (4.3)

of thinking about this overriding process is that the lexical constituent that overrides is actually an elaboration of the eventuality it is attached to, and thus can be modeled with the Elaboration relation within SDRT in the usual way (Figure 4.5). A similar overriding effect is possible using parentheses. Consider (4.4) where it is denoted as an afterthought that the proposed fact is actually currently valid as well as being valid in the past. Note the temporal information in this SDRT because of the special temporal effect created with the parentheses.

[3]The uppercase U with a question mark denotes that the plural pronoun is not yet resolved.

$$
\begin{array}{|l|}
\hline
k_0 := \begin{array}{|l|}\hline \quad \text{U v } e_1 \\ \text{U = ?} \\ e_1\text{:run(U)} \\ \hline \end{array} \\
\qquad \Downarrow \\
k_1 := \begin{array}{|l|}\hline \qquad e_2 \\ e_2\text{:sprint(U)} \\ \text{the hill(v)} \\ \text{up(U,v)} \\ \hline \end{array} \\
\text{Elaboration}(k_0,k_1) \\
\hline
\end{array}
$$

Figure 4.5: Revised SDRS for (4.3)

(4.4) **(WSJ)** The whole notion of "creativity" in education was (and is) part of a romantic rebellion against disciplined instruction.

$$
\begin{array}{|l|}
\hline
k_0 := \begin{array}{|l|}\hline \qquad\quad \text{x y s t n} \\ \text{s} \circ \text{t} \\ \text{t} < \text{n} \\ \text{creativity-in-education(x)} \\ \text{disciplined-instruction(y)} \\ \text{s: rebellion-against(x,y)} \\ \hline \end{array} \\
\qquad \Downarrow \\
k_1 := \begin{array}{|l|}\hline \qquad\quad \text{s'} \\ \text{s'} \circ \text{n} \\ \text{s': rebellion-against(x,y)} \\ \hline \end{array} \\
\text{Elaboration}(k_0,k_1) \\
\hline
\end{array}
$$

Figure 4.6: SDRS for (4.4)

Text clauses conjoined by semicolons with a coordinating conjunction before the last one put special emphasis on the last text-clause as in (4.5).

(4.5) (=(3.28) simplified) Books are nice; some books are exciting; and a private letter is illuminating.

When the text is coherent enough, the informational focus of the last text-clause can be further resolved to a possible topic as in (4.6).

(4.6) (Continued by us from (4.5)) Books are nice; some books are exciting; and a private letter is illuminating. Letters of literary figures have always been popular.

A representational device such as an oval box could signify a further possible expansion as in Figure 4.7. In case of a following sentence such as in (4.6), the oval box can

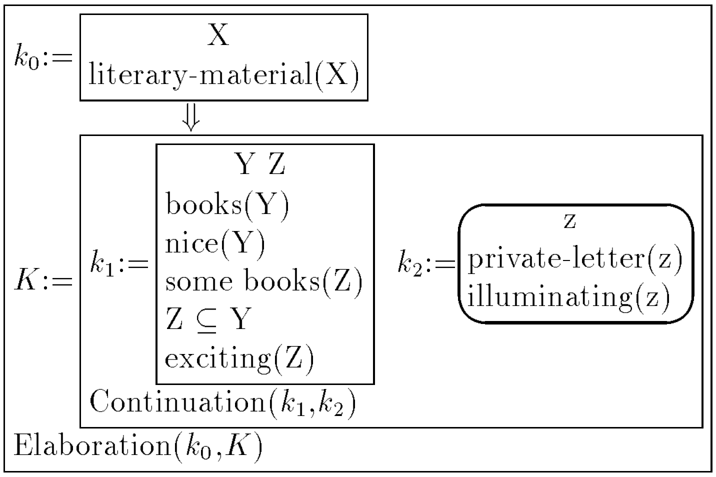

Figure 4.7: SDRS for (4.5)

be dispensed with and the topic mechanism of SDRS can be used as in Figure 4.8.

Let us now consider (4.7) (taken from [Levinson, 1985, p. 134]):

(4.7) a. 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.

b. 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.

Paragraphs (4.7a) and (4.7b) have different interpretations, which are construed in the SDRSs in Figure 4.9. Briefly, (4.7a) cues a subtle degree of explanation where as (4.7b) is more likely to indicate some irony, something more close to Contrast relation by means of the closer coherence provided by the semicolons.

Finally, consider again the (4.8) from Nunberg [1990, p. 31]:

(4.8) a. (=2.3a) He reported the decision: we were forbidden to speak with the chairman directly.

b. (=2.3b) He reported the decision; we were forbidden to speak with the chairman directly.

In (4.8a) there is an Elaboration of the decision itself. In (4.8b) there is an Explanation of the particular manner in which the decision was reported. The particular SDRSs corresponding to these two different relations are shown in Figure 4.10.

4.2.2 Anaphora Resolution

The first modification to deal with anaphora is to encode a way to denote that certain discourse referents are not preferable for selection (though they are available). Both in DRT and SDRT, whether a discourse referent is available as an antecedent is strictly defined with accessibility and availability constraints. However, in sentences such as (4.9), there should be a way to denote that the discourse referents introduced in the dashed sentence are parenthetical and are not preferred for further selection. In (4.9), *He* is resolved to be John, rather than his brother. A similar but stronger effect could also be created by parentheses. Syntactically, a text-clause or a text-sentence enclosed in dashes or parentheses could be a candidate for this effect. A double-framed box is chosen as a representational device for this purpose (Figure 4.11).

(4.9) John—his brother is also an athlete—won the university medal for 3000m easily. He is an ambitious guy.

Normally, both p and q denoting *his brother* and *an athlete* are available for resolution to t as well as j denoting *John* since the SDRS k_2 which includes p and q as discourse referents is rhetorically linked to k_1. The constraint of a parenthetical changes the

k_0:= X
literary-material(X)

$\Downarrow$

K:=

k_1:= Y Z
books(Y)
nice(Y)
some books(Z)
Z $\subseteq$ Y
exciting(Z)

k_2:= Z
private-letters(Z)

$\Downarrow$

k_5:=

k_3:= z
private-letter(z)
illuminating(z)

k_4:= Q W
literary-figures(Q)
Q's private-letters(W)
popular(W)

Continuation(k_3,k_4)

Elaboration(k_2,k_5)

Continuation(k_1,k_2)

Elaboration(k_0,K)

Figure 4.8: SDRS for (4.6)

x u e

k_1:=
y s
Margaret(x)
Gregory(y)
e: meet(x,y)
s o e
Sepik River in NG(u)
s: fall-in-love-at(x,y,u)

k_2:=
z e'
Reo Fortune(z)
x's second husband(z)
e': come-to-work-with-at(x,z,u)

Explanation(k_1, k_2)

x u

k_1:=
y s
Margaret(x)
Gregory(y)
e: meet(x,y)
s o e
Sepik River in NG(u)
s: fall-in-love-at(x,y,u)

k_2:=
z e'
Reo Fortune(z)
x's second husband(z)
e': come-to-work-with-at(x,z,u)

Contrast(k_1, k_2)

Figure 4.9: SDRSs for (4.7a) and (4.7b), respectively

$$\boxed{\begin{array}{l} k_0{:=}\boxed{\begin{array}{l}\qquad\text{x y}\\ \text{decision(x)}\\ \text{reported(y,x)}\\ \text{y=?}\end{array}} \\ \qquad\Downarrow \\ k_1{:=}\boxed{\begin{array}{l}\qquad\qquad\text{U r}\\ \text{U=?}\\ \text{chairman(r)}\\ \text{forbidden-to-speak-directly(U,r)}\end{array}} \\ \text{Elaboration}(k_0,k_1)\end{array}} \qquad \boxed{\begin{array}{l} k_0{:=}\boxed{\begin{array}{l}\qquad\text{x y}\\ \text{decision(x)}\\ \text{reported(y,x)}\\ \text{y=?}\end{array}} \\ \\ k_1{:=}\boxed{\begin{array}{l}\qquad\qquad\text{U r}\\ \text{U=?}\\ \text{chairman(r)}\\ \text{forbidden-to-speak-directly(U,r)}\end{array}} \\ \text{Explanation}(k_0,k_1)\end{array}}$$

Figure 4.10: SDRSs for (4.8a) and (4.8b), respectively

$$\boxed{\begin{array}{l} k_0{:=}\boxed{\begin{array}{l}\quad\text{j}\\ \text{John(j)}\end{array}} \\ \qquad\Downarrow \\ K{:=}\boxed{\begin{array}{l} k_1{:=}\boxed{\begin{array}{l}\qquad\text{y}\\ \text{university-medal(y)}\\ \text{j won y}\end{array}} \quad k_2{:=}\boxed{\boxed{\begin{array}{l}\quad\text{p q}\\ \text{brother(p,q)}\\ \text{athlete(q)}\\ \text{p = j}\end{array}}} \quad k_3{:=}\boxed{\begin{array}{l}\qquad\text{t}\\ \text{ambitious(t)}\\ \text{t = j}\end{array}} \\ \text{Commentary}(k_1,k_2)\\ \text{Continuation}(k_1,k_3)\end{array}} \\ \text{Elaboration}(k_0,K)\end{array}}$$

Figure 4.11: SDRS for (4.9)

preference criteria. Semantically speaking, the underlying semantics of SDRT does not change as the mechanism of the double-framed box will only act as a cue to reorder the available antecedents so that available referents within the SDRS that are double-framed will be moved to less available positions.

For the case of (quoted) speech following colons, the same double-frame device can be used because again the possibility of blocking the outside referents arises. An additional cue would be a reference system change in that indexicals such as the first person pronouns (I/we) would possibly refer to the subject of the preceding text-sentence (Figure 4.12).

(4.10) (Adapted from (3.54)) Lawrence was in Parma, in ecstasy over the work of Correggio: "I went on long visits to the St Jerome and thought about his works."

$$\boxed{\begin{array}{ll} & \text{x y w z U r} \\ k_0 := & \boxed{\begin{array}{l} \qquad\quad \text{e} \\ \text{Lawrence(x)} \\ \text{Parma(y)} \\ \text{e:in(x,y)} \\ \text{Correggio(w)} \\ \text{works(w,U)} \\ \text{in-ecstasy-over(x,U)} \end{array}} \\ k_1 := & \boxed{\boxed{\begin{array}{l} \quad \text{z e' e" O} \\ \text{z=x} \\ \text{StJerome(r)} \\ \text{e':go(z,r)} \\ \text{e":think(z,O)} \\ \text{O=U} \end{array}}} \\ \multicolumn{2}{l}{\text{Explanation}(k_0,k_1)} \end{array}}$$

Figure 4.12: SDRS for (4.10)

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

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

b. 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 will lead to different structures within DRT: in (4.11a) we have *her* attached to Jane and *they* to Joe and Sue, whereas in (4.11b) we have *her* attached to Sue and *they* to Jane and Joe. This difference can be handled with plain DRSs as shown in Figure 4.13.

x y z X Z W
Jane(x)
Joe(y)
Sue(z)
X=y $\oplus$ z
Z=x $\oplus$ X
books-on-England(W)
write(Z,W)
[U | x's books(U), U$\subseteq$W, be(U,best-sellers)] $\Rightarrow$ [be-jealous(X)]

x y z X Z W
Jane(x)
Joe(y)
Sue(z)
X=x $\oplus$ y
Z=X $\oplus$ z
books-on-England(W)
write(Z,W)
[U | z's books(U), U$\subseteq$W, be(U,best-sellers)] $\Rightarrow$ [be-jealous(X)]

Figure 4.13: DRSs for (4.11a) and (4.11b), respectively

As for the effects of restrictive and nonrestrictive clauses, example (4.12a) below implies that Sam has a cat that once belonged to Fred whereas (4.12b) 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, p. 103]). This semantic distinction can straightforwardly be dealt with plain DRSs (cf. Figure 4.14).

(4.12) 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.

$$
\begin{array}{|l|}
\hline
\qquad \text{u Z f s y} \\
\text{Tom(u)} \\
\text{cats(Z)} \\
\text{have(u,Z)} \\
|\text{Z}|=2 \\
\text{Fred(f)} \\
\text{belong(Z,f)} \\
\begin{array}{|l|}
\hline
\quad \text{Y} \\
\text{cats(Y)} \\
\text{have(u,Y)} \\
\text{belong(Y,f)} \\
\hline
\end{array} \Rightarrow \boxed{\text{Y=Z}} \\
\text{Sam(s)} \\
\text{cat(y)} \\
\text{have(s,y)} \\
\begin{array}{|l|}
\hline
\quad \text{t} \\
\text{cat(t)} \\
\text{have(s,t)} \\
\text{belong(t,f)} \\
\hline
\end{array} \Rightarrow \boxed{\text{t=y}} \\
\hline
\end{array}
\qquad\qquad
\begin{array}{|l|}
\hline
\text{u Z f s y} \\
\text{Tom(u)} \\
\text{cats(Z)} \\
\text{have(u,Z)} \\
|\text{Z}|=2 \\
\text{Fred(f)} \\
\text{belong(Z,f)} \\
\text{Sam(s)} \\
\text{cat(y)} \\
\text{have(s,y)} \\
\hline
\end{array}
$$

Figure 4.14: DRSs for (4.12a) and (4.12b), respectively

4.2.3 Presupposition

Punctuation marks such as dashes and parentheses can act as ways to enclose accommodation for a presupposition triggered, for example, by definiteness (see (4.13) where *the black widow of valuation* has a presupposed connotation).

(4.13) [=(3.21) simplified] Dividend growth is "the black widow of valuation"—a reference to the female spiders that attract males and then kill them after mating.

Presupposition has already been examined as a process of specialized anaphora resolution within DRT by Van der Sandt [1992] and in SDRT by Asher and Lascarides [1998]. Van der Sandt takes presuppositions to introduce new DRSs that are marked by a special operator. Then the context of the marked DRS is matched with the previous context. If it does not match, accommodation is realized via adding the required information at a suitable level of accessibility. In Asher and Lascarides [1998], the presupposition has to be rhetorically bound to the context with certain relations such as Background rather than treated as simply anaphoric (as done in Van der Sandt's account). In this thesis, Asher and Lascarides' approach is applied to cases where dashes (or parentheses) mark a constituent that acts as background material for the presupposition they are attached to (see Figure 4.15 for (4.13)).

k_0:=
x y
dividend-growth(x)
black-widow-of-valuation(y)

k_1:=
U V
female-spiders(U)
male-spiders(V)
attract-and-kill(U,V)

Background(k_0,k_1)

Figure 4.15: SDRS for (4.13)

4.3 Other Aspects of the Model

The general method in dealing with effects brought about by punctuation marks is bringing a notational change into (S)DRT (if necessary) and linking that notational change with an existing mechanism within (S)DRT. As to how our simple extensions affect the model-theoretic aspects of DRT and SDRT, amalgamating the proposed changes into the existing mechanisms ensures that the work presented here has the status of an application.

A prototype SDRS builder implementing the model is presented in Appendix C. The chief aim of such an implementation is to show that such implementations are feasible and can provide templates and cues for further use in NLP software such as text summarisation, information extraction modules, and automatic discourse analysers.

4.4 Summary

In this chapter a model of various phenomena signalled by punctuation marks has been presented. What is particular to this approach is the linking of specific syntactic or semantic forms cued by punctuation marks to the information structure of the discourse by using simple extensions to SDRT. There have been applications of SDRT to other natural language phenomena [Asher *et al.*, 1995, Fabricius-Hansen, 1996] but to our knowledge, this is the first such application linking punctuation marks and related occurrences.

Chapter 5

Conclusion

5.1 Contributions

This thesis can best be regarded as a proof of concept; it shows that punctuational devices, though born out of conventions (and are thus prescriptive), do benefit from a descriptive treatment. It supplies considerable evidence to the hypothesis that punctuation could be studied within a text-grammar, as first noticed by Nunberg [1990].

Previous works have supported the latter claim via (computational) contributions emphasising syntax. Here, the outreach is extended by formulating semantic and pragmatic contributions of individual marks to information structure, anaphora resolution, presupposition, discourse relations, and coherence. The semantic and discourse evidence observed and modeled in this way can be considered as a stepping stone towards a unified theory of orthographic devices.

Benefits of constructing a bridge between linguistic observations and semantic modeling are evident. Formal models not only act as a test bed for the linguistic claims made but also help establish engineering-oriented products on a computationally plausible basis. We hope to have adequately shown that our observations correspond to realistic linguistic phenomena that can be modeled in SDRT. Central to our observations is the notion that cues signalled by punctuation marks contribute to the general structure of discourse. Moreover, some of our findings such as the notion of (defeasible) accessibility are novel to DRT itself.

5.2 Limitations and Open Issues

Since obtaining supporting linguistic data has been a prerequisite to carrying out our semantic modeling of punctuation, there is the question of whether idiosyncratic observations have been successfully turned into more general principles in the end. While we realize that the linguistic claims directed toward the four marks studied in this thesis are not equally varied and strong, this is somewhat in the nature of corpus linguistics, and is probably to be expected.

Since we addressed more than one foundational issue in Chapters 3 and 4, it may be claimed that no single issue is examined in great depth. Our aim was not to address an isolated, technical issue in linguistics, but rather, to demonstrate that punctuational cues do contribute to the informational framework within semantics and discourse. “Cueing” as opposed to being “absolute” is a key distinction here: Though punctuation marks cue for linguistic phenomena at semantic or discourse levels, they are still add-ons to the language; thus, they cannot be characterized in an absolute fashion. This does not make them any less valuable to someone who is interested in the information conveyed at the discourse level. It just makes the findings of this thesis to have an element of “defeasibility”, along the lines of some other linguistic phenomena.

Because of the abundance of available data and computational resources, written English has been the object of this study. The lack of a native speaker’s (rather writer’s) intuition on our part is hopefully compensated by our informants and by our decision to work mainly with popular corpora. Accordingly, the major corpus for this study is the WSJ (thus, exhibiting a journalistic style of writing). BC and to a lesser degree, BNC are also used to furnish balance, and to counter the possible effects of style manual adherence or copy-editing on the part of the editors/contributors of the WSJ.

5.3 Future Directions

There is the possibility of strengthening the theoretical contribution of the present work in several directions. Cross-linguistic analyses (for example, a similar study for Turkish) could show how much of our observations are language-specific. Certain non-structural marks and orthographic conventions such as italics and quotations were outside our scope but then again, a similar study on these could bring new insights to the ultimate theory.

There is also scope for experiments assessing how readers incorporate a system of punctuation internally. Apart from Chafe's work [1988], most of the psycholinguistic studies on punctuation have concentrated on teaching punctuation. Additionally, the *comprehension* of punctuated sentences could be analysed by psycholinguistic techniques (e.g., chronometric methods [Caron, 1992, p. 12] which measure the time for comprehension). Testing how the cues presented by punctuation marks are perceived against those cues that are (explicitly) presented otherwise is one such possibility. More studies on the *production* of punctuated sentences could clarify the claims regarding the relationships between intonation and punctuation.

Computationally, the prototype implementation of the SDRT for punctuated sentences carries the seeds of a future module that could be used by other practical modules of NLP, e.g., in text summarization, information extraction, discourse relations detection.

To conclude, semantic and pragmatic information obtained from written sentences come from various sources; punctuation marks are one of them. This corpus-based study has shown some of the approaches for attaining and representing such information by way of a formal model of punctuation.

Appendix A

Discourse Relations

The discourse relations chosen is a core subset of those described by Mann and Thompson [1987]; they are also found in other researchers' work [Asher, 1993, Corston-Oliver, 1998]. Precise definitions of the relations occurring in the following sentences can be found in Asher [1993, pp. 299–309]. The examples below cover all the relations that are shown in the tables of Chapter 3.

Discourse relations occur between two text units, which for the purposes of this thesis, can span units separated or enclosed by punctuation. The main unit without which the pair would be quite incomprehensible is termed the *nucleus* (N) and the supporting unit is termed *satellite* (S). Thus R(N,S) means "the text unit denoted by S is a R (or denotes a relationship of R) to the text unit denoted by N". In the case of symmetric relations between equal-weight nuclei such as Parallel or Contrast, it means "the relation R holds between N and S." Schematically, R(N,S) can be shown as in Figure A.1. The relations are usually characterised according to the eventualities in the text units related. An *eventuality* is the event, state, activity, accomplishment, etc. denoted by the text unit.

1. *Apposition* is a text unit denoting extra information. It may coexist with another relation as does Continuation.

   (BC) Liberals and conservatives in both parties—democratic and republican—shall divorce themselves and form two independent parties, George H. Reama, nationally known labour management expert, said here yesterday.

   (BNC) Of five such founding fathers—Marx, Comte, Spencer, Durkheim and

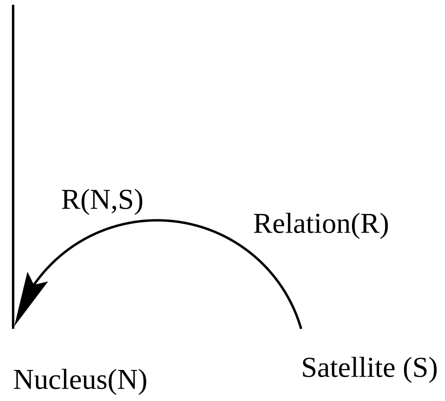

Figure A.1: Schema for Discourse Relations

Weber—Marx (1818–83) and Weber (1864–1920) alone held what could be described as "emancipated" views about women.

2. *Background* holds if the satellite unit is a specialised explanation of the context of the nucleus unit it is attached to .

   (WSJ) Mr. Quinlan, 30 years old, knew he carried a damaged gene, having lost an eye to the rare tumor when he was only two months old—after his mother had suffered the same fate when she was a baby.

3. *Contrast or Parallel* occur between two symmetric units that have some kind of similar structure but exhibit either a contrast or parallel in meaning, respectively.

   (WSJ) Learning skills, producing something cooperatively, feeling useful, they are no longer dependent—others now depend on them.

4. *Continuation* holds between units that have a common topic and where one eventuality is temporally or spatially a continuation of the other. This is the most general relation that can hold between textual parts and may coexist with other relations.

   (BNC) Wallace Arnold (0532–311055) is the accredited coach-tour operator from the UK—a three-day stay at the Hotel Cheyenne for two adults sharing a room ranges from around £130–£150 per person (additional child £65–£81).

5. *Commentary* holds if the satellite text unit is a comment or an opinion stated on the content of the nucleus.

   (WSJ) But as they hurl fireballs that smolder rather than burn, and relive old duels in the sun, it's clear that most are there to make their fans cheer again or recapture the camaraderie of seasons past or prove to themselves and their colleagues that they still have it—or something close to it.

6. *Elaboration* holds if the satellite eventuality is a subpart of and provides details on the eventuality of the nucleus.

   (WSJ) In late trading, the shares were up a whopping 122 pence (\$1.93)—a 16.3% gain—to a record 869 pence on very heavy volume of 9.7 million shares.

   (BC) The social security pay-roll tax is now 6 per cent—3 per cent on each worker and employer—on the first \$4,800 of pay per year.

7. *Explanation* holds if the satellite unit gives supporting reasons, information, etc. on certain aspects of the nucleus unit.

   (BNC) Gary Cattermole remained unbeaten in the latter match, although it was close—defeating Jim Laxton 21–16 in the third, Ron Covall 21–17 in the third and Joe Murray 21–19 in the third.

8. *Instance* holds if the satellite unit is an example used to increase the reader's belief in the topic or the eventuality of the nucleus unit.

   (WSJ) In this connection, it is important to note that several members of New York's sitting City Council represent heterogeneous districts that bring together sizable black, Hispanic, and non-Hispanic white populations—Carolyn Maloney's 8th district in northern Manhattan and the south Bronx and Susan Alter's 25th district in Brooklyn, for example.

9. *Result* holds if the satellite eventuality is the result of the nucleus eventuality happening (thus, if the nucleus eventuality is the *cause* of the satellite eventuality).

   (WSJ) Mr. Steinhardt, who runs about \$1.7 billion for Steinhardt Partners, made his name as a gunslinging trader, moving in and out of stocks with agility—enriching himself and his investment clients.

(BNC) Yesterday, however, American announced that the Stansted–Chicago service will end with the last flight on May 31—putting the jobs of 50 ground staff at risk.

Appendix B

Discourse Representation Theory

B.1 DRT

Discourse Representation Theory (DRT) [Kamp and Reyle, 1993] is a semantic approach that not only shares the concerns of model-theoretic semantics (e.g., Montague grammar) in analysing conditions of truth, but also seeks to relate sentence interpretation to the context in which the sentence is used.

Meanings are assigned by building a series of interpretation structures called Discourse Representation Structures (DRSs) from sentences in the following manner: The initial DRS K_0 represents a starting context for the common ground between the speaker (or the writer) and the audience. The DRS K_1 is then the result of integrating the interpretation of the first sentence to K_0. Continuing in like manner, this process terminates with the construction of K_n which represents the content of the entire discourse [Kamp, 1995].

Clearly, there must be a set of rules which determines on the basis of the syntactic structure of a sentence how the interpretation will be constructed. This is called the DRS construction algorithm. Kamp and Reyle's [1993] original approach involves translation of a parse tree by a number of rewrite rules.[1]

A DRS consists of two parts: A set of *discourse referents* (*discourse markers* or *reference markers*) which represent entities introduced in the discourse and a set of *conditions*

[1] For a summary of different approaches see [Black, 1993]. An elaboration of Asher's [1993] bottom-up technique can be found in Appendix C. For a more recent proposal, see [van Eijck and Kamp, 1997].

that characterize certain properties and relations of those discourse referents. DRSs are constructed by means of certain triggering configurations that activate construction rules. Consider the syntactic tree for (B.1) in Figure B.1.

(B.1) A salesman rings the doorbell.

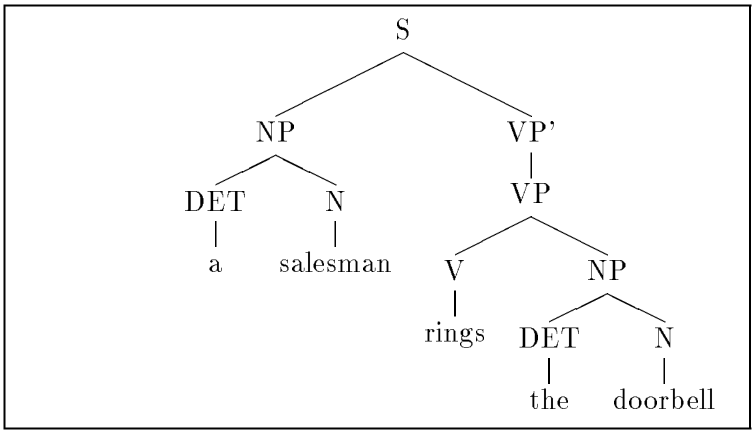

Figure B.1: Syntax tree for (B.1)

Part of the syntax tree in Figure B.1 matches with a triggering configuration as shown in Figure B.2 that will activate the construction rule for Indefinite Descriptions [Kamp and Reyle, 1993, p. 84] stated as follows:

1. Introduce a new discourse referent.
2. Introduce the result of substituting the discourse referent for the NP constituent in the syntactic structure to which the rule is being applied.
3. Introduce a condition in the conditions set to be obtained by placing the discourse referent in parentheses behind the top node of the N constituent.

The resulting DRS from the application of the construction rule for Indefinite Descriptions (the matching part of the template tree in Figure B.2) to the syntax tree shown in Figure B.1 is shown in Figure B.3. Other construction rules further reduce the syntax tree step by step to the DRS for (B.1) as shown in Figure B.4. In the figure, the discourse

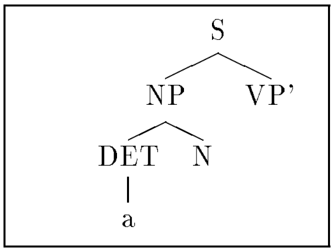

Figure B.2: Triggering Configuration for Indefinite Descriptions

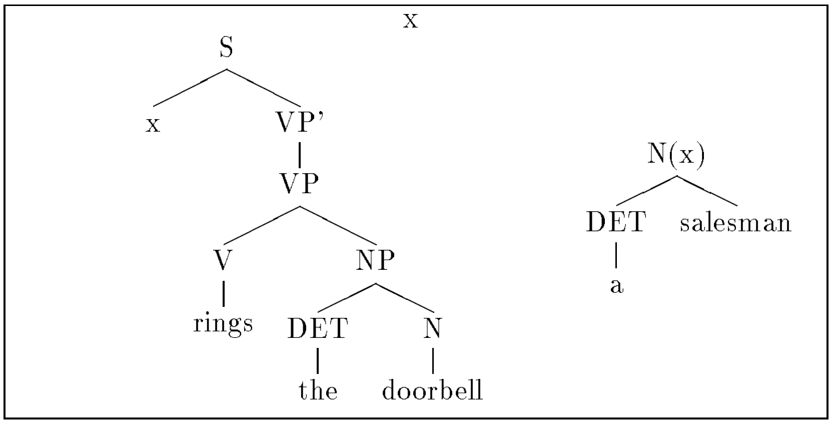

Figure B.3: Interim DRS for (B.1)

referent n represents time of utterance; x, some salesman; z, the doorbell in question; e, the event of the salesman ringing the doorbell. The sign $\subseteq$ denotes that the event occurs as part of time of utterance. (The symbol $\subseteq$ is also used for the quantifier *some* as in "Some As are Bs". The sign $<$ would have denoted an occurrence in the past.)

Let us assume that (B.1) is followed by (B.2).

(B.2) He is selling brushes.

This causes an expansion of the previous DRS as shown in Figure B.5.

It is necessary to link the discourse referents introduced by (B.2) to the context: the discourse referent y stands for the same entity as x (anaphora resolution) and the state of selling brushes overlaps (shown with the sign $\circ$) the event of the ringing of the doorbell. (Brushes are represented with an uppercase U to mark plurality.)

n e x z
e $\subseteq$ n
salesman(x)
doorbell(z)
e: x ring z

Figure B.4: DRS for (B.1)

n e x z y U s
e $\subseteq$ n
salesman(x)
doorbell(z)
e:x ring z
y=x
brushes(U)
s: x selling U
s $\circ$ n
e $\subseteq$ s

Figure B.5: DRS of (B.1) extended with (B.2)

The truth conditions for these kinds of discourse segments are derived from the correspondences of equivalent objects and relations in a model. In DRT terms, a *proper* DRS K (with all discourse referents belonging to the universe of referents of K) is true if and only if there is an *embedding function f* that maps every discourse referent introduced in the universe of K with corresponding elements from the model and *verifies* that the corresponding relationships between the model counterparts hold when there are conditions that relate the original entities in K.

One important aspect of DRT for anaphora resolution is the *accessibility* conditions which constrain how the structure of a DRS affects the resolution of anaphora. Anaphoric constructs must be identified with an accessible discourse referent. DRS A (and thus its referents) are accessible from DRS B if A equals B or A subordinates B. Subordination has several constraints: e.g., A subordinates B if A includes a condition of the form

$A \Rightarrow B$, where the symbol $\Rightarrow$ is used for quantifiers like every and if clauses. Thus, a sentence such as (B.3) is not licensed since the structure introduced by the quantifier *every* results in y not being accessible to z through subordination relation, cf. Figure B.6.

(B.3) Every man loves a woman. She is a beauty.

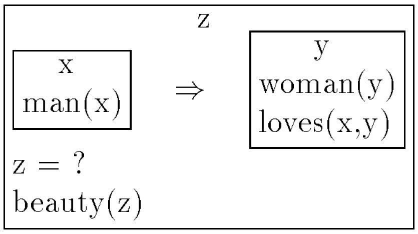

Figure B.6: DRS for (B.3)

DRT has found natural applications in computational linguistics. Two prominent examples include Rossdeutscher [1994] who applies DRT to a German short story, and the Verbmobil speech-to-speech translation system which uses a compositional variant of DRT [Bos *et al.*, 1996].

B.2 SDRT

The original DRT does not consider discourse structure because the DRS for a discourse segment is constructed as one big DRS with a suitable union operation. Segmented Discourse Representation Theory (SDRT) is an extension to DRT for better accounting of systematic effects of discourse structure on discourse interpretation such as abstract entity anaphora [Asher, 1993].

The representational structures of SDRT are called *segmented DRSs* (SDRSs) which have a similar but more complex truth model than that of DRSs.

An SDRS is a pair of sets. The first set contains DRS or SDRSs, while the second pair consists of a set of conditions of discourse relations having the members of the first set as arguments. In SDRT, each new DRS (default unit for a DRS can be a sentence or a clause) has to find a suitable point of attachment. Once that attachment is found, the new structure affects anaphora resolution via constraining discourse structure.

Linking clauses with discourse relations form an important part of SDRT. Asher divides the set of relations he uses into two. Rhetorical relations such as Elaboration, Explanation, and Instance are defined over propositions. Coherence relations such as Cause require a world knowledge component and contribute directly to the truth-conditional content. Two special relations, Parallel and Contrast, license the building of embedding trees to pair the constituents according to similarity/contrast measures. Relations such as Elaboration and Continuation are *topic-based* where a topic *dominates* and summarises the other constituents in the SDRS. Nonmonotonically inferring discourse relations is in some cases made possible by cue words (such as *because, but*) or temporal relations between eventualities. Lascarides and Asher [1993], for example, identify certain temporal relations by means of a defeasible inference mechanism.

As an example of a SDRS for (B.4) where labels only are given for DRSs, see the corresponding SDRS in Figure B.7. Note that each sentence corresponds to a DRS label; a downarrow ($\Downarrow$) indicates topic. In an expression such as $R(K_1, K_2)$ the latter DRS or SDRS stands in a relationship denoted with discourse relation R to the former DRS or SDRS.

(B.4) Kathy was happy to be a teacher. She loved to work with children. Her own mother was a teacher. She still found it a hard job, though.

$$\boxed{\begin{array}{l} \qquad\quad K_1 \\ \qquad\quad \Downarrow \\ K' : \boxed{\begin{array}{l} K_2\ K_3 \\ \text{Continuation}(K_2,\ K_3) \end{array}} \\ \text{Elaboration}(K_1,\ K') \\ K_4 \\ \text{Contrast}(K_1,\ K_4) \end{array}}$$

Figure B.7: SDRS for (B.4)

In Figure B.7, the topic DRS dominating the rest allows *it* (a case of abstract entity anaphora) in the last sentence to be resolved to *being a teacher* by means of attachment criteria of SDRT. A question arises as to which constituents serve as attachment points for new units. This point requires the definition of discourse-subordination (d-subordination). A SDRS α is d-subordinate to β if and only if α is a constituent of β transitively or β is a topic for α (denoted as $\beta \Downarrow \alpha$) or α is declared in the universe of β. Available attachment sites for the new unit are either the current constituent or the SDRS to which the current constituent is d-subordinate. Topic-based updating with relations such as Elaboration and Continuation require this criteria where non-topic based relations such as Cause, Contrast, etc. also require the attachment point to be d-free (i.e., not d-subordinate to a further SDRS). Apart from the normal accessibility constraints of DRT, SDRT also enforces a number of *availability* constraints which constrain possible antecedents of an anaphor to a discourse related constituent or the current constituent or their discourse referents or subDRSs. (See [Asher, 1993, Ch. 7] for details.)

Among applications of SDRT (such as [Asher *et al.*, 1995]) Fabricius-Hansen's [1996] is worth mentioning as she strives to use SDRT to show the effects of some linguistic phenomena in translation theory in a similar way attempted in this thesis. The difficulties raised when translating from a syntactically complex and informationally dense language (German) into a less complex one (Norwegian) are explored by her by means of several measures applied into constructing SDRSs. Two such measures are information splitting (such as having to split sentences of the source language into more sentences and clauses of the target one) and discourse structure fidelity (the degree of conformance to the source language's discourse structure).

Appendix C

A Simple Prototype

C.1 An SDRT Prototype

The main requirement for the SDRT prototype would be to allow the user to enter a number of possibly punctuated sentences and be presented with an SDRS where punctuation marks affect the discourse in a number of places. To this end, a parser and a grammar for English are needed to deal with punctuated sentences syntactically. The input sentences would then be fed into the parser one by one, the first sentence acting as an initial context for the following. The remaining sentences would be furnished with their own (S)DRSs and be incrementally added to the main SDRS as dictated by the kind of discourse relation detected. This main SDRS could later be fed into other software modules as required. The basic usage of the prototype is shown in Figure C.1.

As explained and modeled in Chapters 3 and 4, certain referents override the default accessibility mechanisms of (S)DRT, either by being made more prominent (such as an NP following a colon at the end of a sentence) or less prominent than they would otherwise be (such as a referent introduced in a sentence within parentheses). Consequently, there should be appropriate mechanisms to cater for the overriding of default accessibility. Similarly, if a certain syntactic pattern of a punctuation mark cues for a certain discourse relation, that discourse relation should defeasibly be made the discourse relation for structuring the discourse.

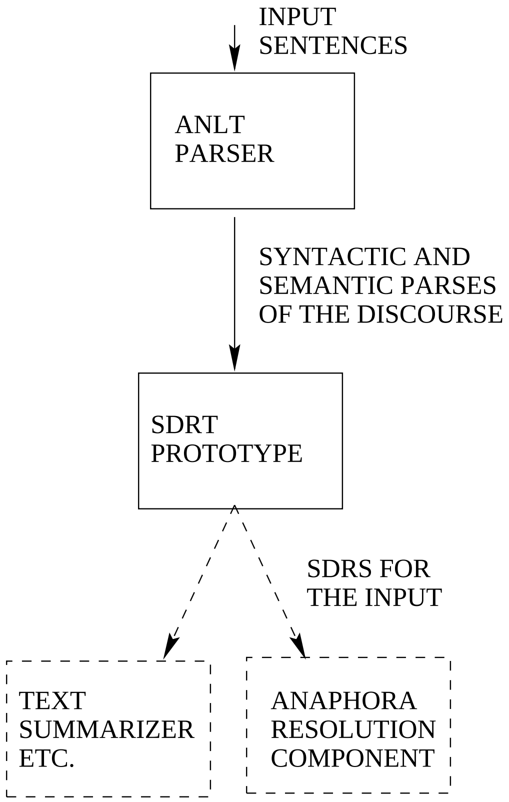

Figure C.1: Basic Usage of the Prototype

C.2 Design Strategy

The requirement of having a non-trivial, wide-coverage parser for English that would also handle punctuated sentences is fulfilled by using Alvey Natural Language Tools (ANLT) analyser and grammar developed at Cambridge University, UK. ANLT grammar is an English grammar based on a formalism similar to the Generalized Phrase Structure Grammar (GPSG) [Gazdar *et al.*, 1985]. It comprises a general-purpose, wide-coverage, sentence-based morphosyntactic and semantic analyser for English [Grover *et al.*, 1993]. Different rule types exist for the syntactic analysis of a given sentence combined with an event-based and λ-calculus based compositional semantics. No provision is made for discourse-level phenomena or anaphora resolution in the ANLT grammar itself. This grammar already had syntactic rules dealing with punctuation as implemented by Briscoe [1994, 1996]. A partial implementation of semantic rules distinguishing coordination and subordination was done by Lee [1995]. To the semantic rules, we made some modifications for integrating the effects of individual punctuation marks.

There are several ways to build DRSs computationally. A top-down approach triggered by syntactic patterns is presented in Kamp and Reyle [1993]. Bottom-up (and compositional) approaches that are lexically driven are also available, such as the λ-DRT [Bos *et al.*, 1996]. A similar bottom-up approach [Wada and Asher, 1986, Asher, 1993] is adopted in this thesis.

Input is assumed to consist of the syntactic parses of sentences in the original approach of Asher [1993, pp. 69–75]. This thesis uses the semantic output. Lexical items have associated *partial* or *predicative* DRSs. Common nouns and verbs yield predicative DRSs where discourse referents are allowed to have placeholders by means of the λ operator and variables. For example, for the common noun "abbot" the corresponding predicative DRS will be of the form $\lambda x \cdot abbot(x)$ where the lambda expression will later be applied on a discourse referent. Pictorially, this can be shown as in Fig C.2.

$$\lambda\mathrm{x}\,\boxed{\mathrm{abbot(x)}}$$

Figure C.2: Predicative DRS for "abbot"

Determiners and specifiers, on the other hand, introduce partial DRSs which may include declared discourse referents with predicative DRSs abstracted away. For example, the determiner "a" will have the partial DRS in Figure C.3 embodying placeholders for the partial DRSs. A proper name such as "Kim" also introduces a partial DRS with

$$\lambda\mathrm{P}\lambda\mathrm{Q}\,\boxed{\begin{array}{c}\mathrm{u}\\ \mathrm{P(u)}\\ \mathrm{Q(u)}\end{array}}$$

Figure C.3: Partial DRS for "a"

a slot already filled in, see Figure C.4. By means of *DRS conversion*, partial and predicative DRSs are merged incrementally. The noun phrase "an abbot" will then get

$$\lambda P \begin{array}{|c|}\hline \text{u} \\ \text{Kim(u)} \\ \text{P(u)} \\ \hline\end{array}$$

Figure C.4: Partial DRS for “Kim”

$$\lambda Q \begin{array}{|c|}\hline \text{u} \\ \text{abbot(u)} \\ \text{Q(u)} \\ \hline\end{array}$$

Figure C.5: Partial DRS for “an abbot”

a partial DRS as in Figure C.5. This conversion process will start from the object of the sentence and continue right-to-left till the subject is processed and a full DRS which can then be passed onto an anaphora processing module is produced .

In core DRS, additional DRSs will also be merged with this initial DRS to form one main DRS for the discourse. Since discourse structure is relevant for us, a different procedure is used in this thesis for combining DRSs into SDRSs noting the discourse relations between them and their structure. Parenthetical referents are held in a stack once they are captured by means of punctuational cues. The resulting main SDRS of a series of sentences has the structure shown in Figure C.6. *Label* is used for labeling the SDRSs and DRSs. If the *Discourse Relation* involved requires a topic, then a *Topic* DRS is created. Discourse relations relate two subDRSs or SDRSs whereas the Parenthetical stack reflects the defeasible accessibility criteria and reorders the accessible referents.

The implementation language is Common LISP [Graham, 1996]. The prototype works as a stand-alone module with a simple user interface that allows the usage and the selection of previously processed sentences, viz. the output of the ANLT analyser. After it is constructed, the SDRS output can be viewed and saved. Detailed information on the functional operation of the system is given in the next section.

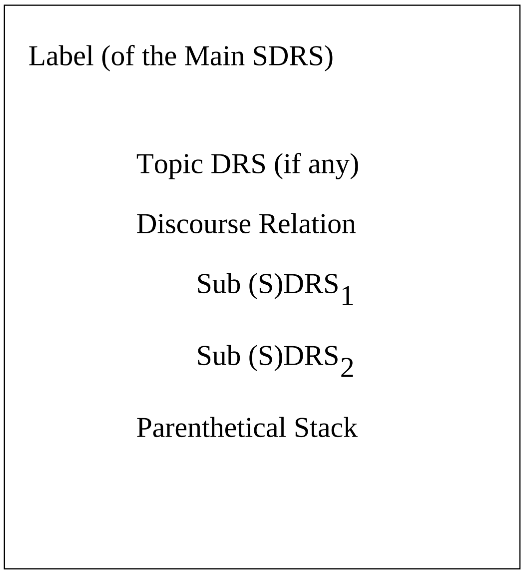

Figure C.6: Structure of the Main SDRS

C.3 Implementation

C.3.1 Functional Description

Currently, the SDRT prototype works as postprocessor to the ANLT analyser. Functional structure of the prototype can be explained in three main steps along with examples.

1. **Preprocessing the Input:** Once the user chooses the set of input sentences to be processed, the relevant semantic forms of the sentences chosen are taken from the preexisting input file (See Figure C.7 for semantic form of the sentence "Kim—Lee is crazy—will abdicate"). These semantic forms are outputs of the ANLT analyser. Relevant accessor functions are designed according to the format of these semantic forms to access the lexical forms associated with partial and predicative DRSs in the lexicon. The discourse referents already generated by ANLT (unique within a single text-sentence only) are preprocessed to ensure uniqueness throughout the discourse.

2. **Building a Discourse Structure:** A main SDRS is built incrementally by using the SDRS building strategy detailed in Section C.2. After the initial SDRS (which

```
(DASH
 (DECL
  (ABDICATE (uqe (some (e1) (FUT e1)))
   (name (the (x1) (and (sg x1) (named x1 KIM) (animate x1)))))
  (BE (uqe (some (e2) (PRES e2)))
   (CRAZY (name (the (x2) (and (sg x2) (named x2 LEE) (animate x1))))
  (degree unknown)))))
```

Figure C.7: Input Semantic Form

can be indeed a DRS), the structure is built on top of the existing context according to the structure of the sentence, in particular according to the punctuation marks encountered.

Let us examine the corresponding outputs for three punctuated sentences.

(C.1) Kim—Lee is crazy—will abdicate.

(C.2) She gives him a message: two weeks.

(C.3) An abbot helped—or promised to help—Kim.

The following phenomena are explained in greater length in Chapters 3 and 4. Figure C.8 shows the SDRS built for (C.1). The discourse referent that denotes `LEE` is put at the bottom of the stack `PAREN?` as it is going to be defeasibly inaccessible. The SDRS in Figure C.9 for (C.2) has the NP following the colon as a topic of the sentence. The SDRS in Figure C.10 for (C.3) handles the case of dashes acting cues for one eventuality being overridden by another by means of Elaboration relation. Once the SDRS is built it is saved in a file.

C.3.2 Assumptions, Constraints, and Integration

One assumption is that the output of the ANLT parser is not ambiguous. This is not a realistic assumption as the output for real sentences do contain many parses. Ambiguous parses produced by the ANLT parser are eliminated manually.

Other constraints relate to the scope of the implementation. DRT deals with several linguistic phenomena such as quantifiers, tense, and aspect. Since the point in building

```
SR-140
 CONTINUATION

  DR-141
   X10

   (KIM X10)
   (ABDICATE X10)

  DR-142

   X20

   (LEE X20)
   (CRAZY X20)
  PAREN? (X20)
```

Figure C.8: Output of the prototype for (C.1)

this prototype implementation is not to develop a full-fledged SDRS builder but to show that one can take advantage of the cues provided by punctuation marks, this version of the implementation does not deal with several such phenomena. Similarly, a separate core lexicon has been used instead of the full vocabulary of ANLT grammar. Moreover, due to hardships in getting the ANLT parser to produce the right semantic forms, some of the forms were manipulated by the author.

By the same token, there needs to be a discourse relation discovery module that makes use of several factors of context (for example, cue words, eventual aspects) in addition to taking discourse relations cued by punctuation marks. Currently, apart from the relations cued by the punctuation mark (see Chapter 3) the default relation "Continuation" is used. Similarly, an anaphora resolution module is not implemented.

```
SR-120
 TOPIC
  DR-121
    X40

    (TWO X40)
    (WEEKS X40)

 ELABORATION

  DR-122
   X10 X20 X30

   (?  X10)
   (MESSAGE X20)
   (GIVES X10 X30 X20)
   (? X30)
```

Figure C.9: Output of the prototype for (C.2)

```
SR-123
 TOPIC
  X10 X20

  (ABBOT X10)
  (KIM X20)

 ELABORATION

 DR-130

  (HELP  X10 X20)

 DR-131

  (PROMISE HELP  X10 X20)
```

Figure C.10: Output of the prototype for (C.3)

Bibliography

[Akram and Saadeddin, 1987] M. Akram and A. M. Saadeddin. Target-World Experiential Matching: The Case of Arabic/English Translating. *Quinquereme*, 10(2):137–164, 1987.

[Asher, 1993] N. Asher. *Reference to Abstract Objects in Discourse.* Kluwer, Dordrecht, Netherlands, 1993.

[Asher and Lascarides, 1998] N. Asher and A. Lascarides. The Semantics and Pragmatics of Presupposition, *Journal of Semantics*, 1998 (Forthcoming).

[Asher *et al.*, 1995] N. Asher, M. Aurnague, M. Bras, and L. Vieu. Spatial, Temporal and Spatio-temporal Locating Adverbials in Discourse. In *Time, Space, and Movement (TSM)*, pages 107–119, Gascony, France, 1995.

[Bayraktar *et al.*, 1998] M. Bayraktar, B. Say, and V. Akman. An Analysis of English Punctuation: The Special Case of Comma. *International Journal of Corpus Linguistics*, 3(1):33–57, 1998.

[Beaver, 1997] D. I. Beaver. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 939–1008. Elsevier, Amsterdam, Netherlands, 1997.

[Black, 1993] A. W. Black. Some Different Approaches to DRT. In Robin Cooper, editor, *Integrating Semantic Theories*, Dyana-2 (Dynamic Interpretation of Natural Language) Report, R 2.1.A, pages 101–120. 1993.

[Blakemore, 1996] D. Blakemore. Are Apposition Markers Discourse Markers? *Journal of Linguistics*, 32:325–347, 1996.

[BNC, 1997] BNC. (British National Corpus). Information available on the WWW: http://info.ox.ac.uk/bnc/

[Bolinger, 1989] D. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford University Press, Stanford, CA, 1989.

[Bos *et al.*, 1996] J. Bos, B. Gambäck, C. Lieske, Y. Mori, M. Pinkal, and K. Worm. Compositional Semantics in Verbmobil. In *Proceedings of 16th International Conference on Computational Linguistics (COLING '96)*, pages 131–136, Copenhagen, Denmark, 1996.

[Briscoe, 1994] T. Briscoe. Parsing (with) Punctuation. Technical Report, Rank Xerox Research Centre, Grenoble, France, 1994.

[Briscoe, 1996] T. Briscoe. The Syntax and Semantics of Punctuation and Its Use in Interpretation. In [Jones, 1996a], pages 1–8.

[Briscoe and Carroll, 1995] T. Briscoe and J. Carroll. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of International Workshop on Parsing Technologies*, pages 48–58, Prague, Czech Republic, 1995.

[Caron, 1992] J. Caron. *An Introduction to Psycholinguistics*. University of Toronto Press, Toronto, Canada, 1992.

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

[Cormack, 1992] S. Cormack. *Focus and Discourse Representation Theory*. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1992.

[Corston-Oliver, 1998] S. H. Corston-Oliver. Computing Representations of the Structure of Written Discourse. Technical Report MSR-TR-98-75, Microsoft Research Institute, Redmond, WA, 1998.

[Cruttenden, 1986] A. Cruttenden. *Intonation.* Cambridge University Press, Cambridge, UK, 1986.

[Dale, 1991a] R. Dale. 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*, pages 110–120. Technical University of Berlin, Berlin, Germany, 1991.

[Dale, 1991b] R. Dale. The Role of Punctuation in Discourse Structure. In *Working Notes for AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, pages 13–14. Asilomar, CA, 1991.

[de Beaugrande and Dressler, 1986] R. de Beaugrande and W. Dressler. *Introduction to Text Linguistics.* Longman, London, 1986.

[Doran, 1996] C. Doran. Punctuation in Quoted Speech. In [Jones, 1996a], pages 9–18.

[Ehrlich, 1992] E. Ehrlich. *Theory and Problems of Punctuation, Capitalization, and Spelling.* McGraw-Hill, Hong Kong, 1992.

[Engdahl and Vallduví, 1996] E. Engdahl and E. Vallduví. The Linguistic Realization of Information Packaging. *Linguistics*, 34:459–519, 1996.

[Fabricius-Hansen, 1996] C. Fabricius-Hansen. Informational Density: A Problem for Translation and Translation Theory. *Linguistics*, 34:521–565, 1996.

[Francis and Kučera, 1982] W. N. Francis and H. Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston, MA, 1982.

[Garside *et al.*, 1987] R. Garside, G. Leech, and G. Sampson, editors. *The Computational Analysis of English.* Longman, London, 1987.

[Gazdar *et al.*, 1985] G. Gazdar, E. Klein, G. K. Pullum, and I. Sag. *Generalized Phrase Structure Grammar.* Blackwell, Oxford, UK, 1985.

[Graham, 1996] P. Graham. *ANSI Common Lisp.* Prentice Hall, New Jersey, NJ, 1996.

[Grosz and Sidner, 1986] B. J. Grosz and C. L. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[Grosz *et al.*, 1995] B. Grosz, A. Joshi, and S. Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[Grover *et al.*, 1993] C. Grover, J. Carroll, and T. Briscoe. The Alvey Natural Language Tools Grammar. Technical Report 284, Computer Laboratory, Cambridge University, Cambridge, UK, 1993.

[Hadumod, 1996] B. Hadumod. *Routledge Dictionary of Language and Linguistics*. Routledge, London, UK, 1996.

[Hall and Robinson, 1996] N. Hall and A. Robinson. The Punctuation Project, 1996. Available on the WWW: http://bll.edu.aca.mmu.ac.uk/punctuation.html

[Harris, 1995] R. Harris. *Signs of Writing*. Routledge, London, UK, 1995.

[Hendriks, 1996] H. Hendriks. Information Packaging. From Cards to Boxes. In T. Galloway and J. Spence, editors, *Proceedings of Semantics and Linguistics Theory (SALT VI)*, Cornell University, Ithaca, NY, 1996.

[Hovy and Arens, 1991] E. H. Hovy and V. Arens. Automatic Generation of Formatted Text. In *Proceedings of 9th National Conference on Artificial Intelligence (AAAI '91)*, pages 92–96, MIT Press, Cambridge, MA, 1991.

[Humphreys, 1993] L. Humphreys. Book Review: The Linguistics of Punctuation. *Machine Translation*, 7:199–201, 1993.

[Hurst and Douglas, 1997] M. Hurst and S. Douglas. Layout & Language: Preliminary Experiments in Assigning Logical Structure to Table Cells. In *Proceedings of 5th Applied Natural Language Processing Conference*, pages 217–220, Washington, D.C., 1997.

[Jones, 1994a] B. Jones. Can Punctuation Help Parsing? Acquilex-II Working Paper 29, Computer Laboratory, Cambridge University, Cambridge, UK, 1994.

[Jones, 1994b] B. Jones. Exploring the Role of Punctuation in Parsing Natural Language. In *Proceedings of 15th International Conference on Computational Linguistics (COLING '94)*, pages 421–425, Kyoto, Japan, 1994.

[Jones, 1995] B. Jones. Exploring the Variety and Use of Punctuation. In *Proceedings of 17th Annual Cognitive Science Conference*, pages 619–624, Pittsburgh, PA, 1995.

[Jones, 1996a] B. Jones, editor. *Punctuation in Computational Linguistics*, University of California Santa Cruz, Santa Cruz, CA, 1996. *SIGPARSE 1996* (Post Conference Workshop of *ACL96*). Available on the WWW: http://www.cogsci.ed.ac.uk/hcrc/publications/wp-2.html

[Jones, 1996b] B. Jones. Towards a Syntactic Account of Punctuation. In *Proceedings of 16th International Conference on Computational Linguistics (COLING '96)*, pages 604–609, Copenhagen, Denmark, 1996.

[Jones, 1996c] B. Jones. Towards Testing the Syntax of Punctuation. In *Proceedings of 34th Annual Meeting of Association for Computational Linguistics—Student Session*, pages 363–365, Santa Cruz, CA, 1996.

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

[Kamp, 1995] H. Kamp. Discourse Representation Theory. In J. Verschueren, J. O. Östman, and J. Blommaert, editors, *Handbook of Pragmatics: Manual*, pages 253–257. John Benjamins, Amsterdam, Netherlands, 1995.

[Kamp and Reyle, 1993] H. Kamp and U. Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Kluwer, Dordrecht, Netherlands, 1993.

[Karlsson *et al.*, 1994] F. Karlsson, A. Voutilainen, J. Heikkila, and A. Antilla, editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, Germany, 1994.

[Kennedy and Boguraev, 1996] C. Kennedy and B. Boguraev. Anaphora for Everyone: Pronominal Anaphora Resolution Without a Parser. In *Proceedings of 16th International Conference on Computational Linguistics (COLING '96)*, pages 113–118, Copenhagen, Denmark, 1996.

[Kessler *et al.*, 1997] B. Kessler, G. Nunberg, and H. Schütze. Automatic Detection of Text Genre. In *Proceedings of 35th Annual Meeting of Association for Computational Linguistics and 8th Conference of European Chapter of Association for Computational Linguistics*, pages 32–38, Madrid, Spain, 1997.

[Kučera, 1992] H. Kučera. Brown Corpus. In S. C. Shapiro, editor, *The Encyclopedia of Artificial Intelligence*, pages 128–130. John Wiley, New York, second edition, 1992.

[Lascarides and Asher, 1993] A. Lascarides and N. Asher. Temporal Interpretation, Discourse Relations, and Commonsense Entailment. *Linguistics and Philosophy*, 16:437–493, 1993.

[Lavoie and Ranbow, 1997] B. Lavoie and O. Ranbow. A Fast and Portable Realizer for Text Generation Systems. In *Proceedings of 5th Conference on Applied Natural Processing*, pages 265–268, Washington, D.C., 1997.

[Lee, 1995] S. Lee. A Syntax and Semantics for Text Grammar. Master's thesis, Engineering Department, Cambridge University, Cambridge, UK, 1995.

[Levinson, 1985] J. P. Levinson. *Punctuation and the Orthographic Sentence: A Linguistic Analysis.* Ph.D. thesis, Department of Linguistics, City University of New York, 1985.

[Lewis, 1979] D. Lewis. Scorekeeping in a Language Game. *Journal of Philosophical Logic*, 8:339–359, 1979.

[Mann and Thompson, 1987] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report RS-87-190, Information Sciences Institute, University of Southern California, Marina Del Rey, CA, 1987.

[Marcu, 1997] D. Marcu. The Rhetorical Parsing of Natural Language Texts. In *Proceedings of 35th Annual Meeting of Association for Computational Linguistics and 8th Conference of European Chapter of Association for Computational Linguistics*, pages 96–103, Madrid, Spain, 1997.

[Marcus *et al.*, 1993] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

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

[McDermott, 1990] J. McDermott. *Punctuation for Now.* MacMillan, Hong Kong, 1990.

[Meyer, 1986] C. F. Meyer. Punctuation Practice in the Brown Corpus. *ICAME Newsletter*, pages 80–95, 1986.

[Meyer, 1987] C. F. Meyer. *A Linguistic Study of American Punctuation.* Peter Lang, New York, 1987.

[Min, 1996] Y. G. Min. Role of Punctuation in Disambiguation of Coordinate Compounds. In [Jones, 1996a], pages 33–40.

[Mitkov and Boguraev, 1997] R. Mitkov and B. Boguraev, editors. *Operational Factors in Practical Anaphora Resolution for Unrestricted Texts* (Postconference Workshop of ACL/EACL '97), Madrid, Spain, 1997.

[Nunberg, 1990] G. Nunberg. *The Linguistics of Punctuation.* CSLI Publications, Stanford, CA, 1990.

[Nunberg, 1997] G. Nunberg. New Frontiers in Punctuation Research. Linguistics Department Colloquium, Stanford University, CA, 2 May 1997.

[Osborne, 1996] M. Osborne. Can Punctuation Help Learning? In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 399–412. Springer-Verlag, Berlin, Germany, 1996.

[Palmer and Hearst, 1994] D. Palmer and M. Hearst. Adaptive Sentence Boundary Disambiguation. In *Proceedings of Applied Natural Language Processing Conference (ANLP '94)*, pages 78–83, Stuttgart, Germany, 1994.

[Parkes, 1993] M. B. Parkes. *Pause and Effect: An Introduction to the History of Punctuation in the West.* University of California Press, Berkeley, CA, 1993.

[Partridge, 1953] E. Partridge. *You Have a Point There: A Guide to Punctuation and its Allies.* Hamish Hamilton, London, 1953. Reprinted by Routledge in 1993.

[Pascual and Virbel, 1996] E. Pascual and J. Virbel. Semantic and Layout Properties of Text Punctuation. In [Jones, 1996a], pages 41–47.

[Pascual, 1996] E. Pascual. Integrating Text Formatting and Text Generation. In G. Adorni and M. Zock, editors, *Trends in Natural Language Generation: An Artificial Intelligence Perspective*, pages 205–221. Springer-Verlag, Berlin, Germany, 1996.

[Pullum, 1991] G. K. Pullum. Punctuation and Human Freedom. In *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language*, University of Chicago Press, Chicago, IL, 1991.

[Quirk *et al.*, 1972] R. Quirk, S. Greenbaum, G. Leech and J. Svartvik. *A Grammar of Contemporary English.* Longman, London, UK, 1972.

[Reed and Long, 1997] C. Reed and D. Long. Generating Punctuation in Written Arguments. Technical Report RN/97/157, Department of Computer Science, University College, London, UK, 1997.

[Reynar and Ratnaparkhi, 1997] J C. Reynar and A. Ratnaparkhi. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of 5th Applied Language Processing Conference*, pages 16–19, Washington, D.C., 1997.

[Robinson, 1988] G G. Robinson. The Punctuator's World: A Discursion. *Syracuse University Library Associates Courier*, XXIII(2):73–104, 1988.

[Robinson, 1989] G G. Robinson. The Punctuator's World: A Discursion. Part 2. The Grammarians: AD 250–AD 1250. *Syracuse University Library Associates Courier*, XXIV(1):63–99, 1989.

[Robinson, 1990a] G G. Robinson. The Punctuator's World: A Discursion. Part 4. England Upto 1650. *Syracuse University Library Associates Courier*, XXV(1):85–125, 1990.

[Robinson, 1990b] G. G. Robinson. The Punctuator's World: A Discursion. Part 5. Logic Takes Over: 1650–1775. *Syracuse University Library Associates Courier*, XXV(2):81–121, 1990.

[Robinson, 1992] G. G. Robinson. The Punctuator's World: A Discursion. Part 7. Age of Pragmatism: 1800–1850. *Syracuse University Library Associates Courier*, XXVII(1):111–158, 1992.

[Robinson, 1996] G. G. Robinson. The Punctuator's World: A Discursion. Part 9. Stirrings of Retreat: 1900 to Midcentury. *Syracuse University Library Associates Courier*, XXXI:75–106, 1996.

[Robinson, 1997] G. G. Robinson. The Punctuator's World: A Discursion. Part 10. 1950: Onwards! But where? *Syracuse University Library Associates Courier*, XXXII:123–151, 1998.

[Rossdeutscher, 1994] A. Rossdeutscher. Fat Child Meets DRT: A Semantic Representation for the Opening Lines of Kaschnitz' "Das Dicke Kind". *Theoretical Linguistics*, 20(2/3):238–304, 1994.

[Sampson, 1992] G. Sampson. Book Review: The Linguistics of Punctuation. *Linguistics*, 30(2):467–475, 1992.

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

[Say, 1997] B. Say. An Information-Based Approach to Punctuation. In *Proceedings of 14th National Conference on Artificial Intelligence (AAAI '97)*, page 818, AAAI Press, Menlo Park, CA, 1997.

[Say and Akman, 1996] B. Say and V. Akman. Information-Based Aspects of Punctuation. In [Jones, 1996a], pages 49–56.

[Say and Akman, 1997] B. Say and V. Akman. Current Approaches to Punctuation in Computational Linguistics. *Computers and the Humanities*, 30(6):457–469, 1997.

[Say and Akman, 1998a] B. Say and V. Akman. An Information-Based Treatment of Punctuation in Discourse Representation Theory. In C. Martín-Vide, editor, *Mathematical and Computational Analysis of Natural Language*, pages 359–373, John Benjamins, Philadelphia, PA, 1998.

[Say and Akman, 1998b] B. Say and V. Akman. Dashes as Typographical Cues for Information Structure. In *3rd Conference in Information-Theoretic Approaches to Logic, Language, and Computation (ITALLC '98)*, pages 209–223, Hsitou, Taiwan, 1998.

[Schiffrin, 1987] D. Schiffrin. *Discourse Markers*. Cambridge University Press, Cambridge, UK, 1987.

[Scholes and Willis, 1990] R. J. Scholes and B. J. Willis. Prosodic and Syntactic Functions of Punctuation—A Contribution to the Study of Orality and Literacy. *Interchange*, 21(3):13–20, 1990.

[Seuren, 1994] P. Seuren. Accommodation and Presupposition. In R. E. Asher, editor, *The Encyclopedia of Language and Linguistics*, pages 15–16. Pergamon Press, Oxford, UK, 1994.

[Shiuan and Ann, 1996] P. L. Shiuan and C. Ting Hian Ann. A Divide-and-Conquer Strategy for Parsing. In [Jones, 1996a], pages 57–66.

[Simard, 1996] M. Simard. Considerations on Parsing a Poorly Punctuated Text in French. In [Jones, 1996a], pages 67–72.

[Smith, 1986] C. L. Smith. Attitudinal Study of Graphic Computer-Based Instruction for Punctuation. *Journal of Technical Writing and Communication*, 3:267–272, 1986.

[Sperber and Wilson, 1986] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, UK, 1986.

[Taylor and Knowles, 1988] L. J. Taylor and G. Knowles. Manual of Information to Accompany the SEC Corpus. University of Lancaster, Lancaster, UK, 1988.

[Turan, 1997] Ü. D. Turan. Ranking Forward-Looking Centers in Turkish: Universal and Language-Specific Properties. In M. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*, pages 139–160. Oxford University Press, Oxford, UK, 1997.

[Twine, 1984] N. Twine. The Adoption of Punctuation in Japanese Script. *Visible Language*, 18(3):229–237, 1984.

[Vallduví, 1992] E. Vallduví. *The Informational Component*. Garland, New York, 1992.

[Vallduví, 1993] E. Vallduví. Information packaging: A Survey. Technical report, No: HCRC/RP-44. Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1993. Available on the WWW: http://www.cogsci.ed.ac.uk/hcrc/publications/

[van der Sandt, 1992] R. van der Sandt. Presupposition Projection as Anaphora Resolution. *Journal of Semantics*, 9:333–377, 1992.

[van Eijck and Kamp, 1997] J. van Eijck and H. Kamp. Representing Discourse in Context. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 179–237. Elsevier, Amsterdam, Netherlands, 1997.

[Wada and Asher, 1986] H. Wada and N. Asher. BUILDRS: An Implementation of DR Theory and LFG. In *Proceedings of 11th International Conference on Computational Linguistics (COLING '86)*, pages 540–545, University of Bonn, Bonn, Germany, 1986.

[White, 1995] M. White. Presenting Punctuation. In *Proceedings of 5th European Workshop on Natural Language Generation*, pages 107–125, Leiden, Netherlands, 1995.

Index