What's The Point?

A (Computational) Theory of Punctuation

Bernard Jones

PhD
The University of Edinburgh
1996

Abstract

Although punctuation is clearly an important part of the written language, many natural language processing systems developed to date simply ignore punctuation in input text, or do not place it in output text. The reason for this is the lack of any clear, implementable theory of punctuation function suitable for transfer to the computational domain.

The work described in this thesis aims to build on previous linguistic work on the function of punctuation, particularly that by Nunberg (1990), with experimental and theoretical investigations into the potential usefulness of including punctuation in natural language analyses, the variety of punctuation marks present in text, and the syntactic and semantic functions of those marks . Results from these investigations are combined into a taxonomy of punctuation marks and synthesised into a theory describing principles and rule schemata whereby punctuation functionality can be added to natural language processing systems.

The thesis begins with some introductory chapters, discussing the nature of punctuation, its history, and previous approaches to theoretical description. Subsequent chapters describe the experimental and theoretical investigations into the potential uses of punctuation in computational systems, the variety of punctuation marks used, and the syntactic and semantic functions that punctuation marks fulfil. Further chapters then construct a taxonomy of punctuation marks and describe the theory synthesised from the results of the investigations. The concluding chapters sum up the research and discuss its possible extension to languages other than English.

"You must please remember that a theme, a thesis, is in most cases little more than a sort of clothes line on which one pegs a string of ideas, quotations, allusions and so on, one's mental undergarments of all shapes and sizes, some possibly fairly new but most rather old and patched; and they dance and sway in the breeze and flap and flutter, or hang limp and lifeless; and some are ordinary enough, and some are of a private and intimate shape, and rather give the owner away, and show up his or her peculiarities. And owing to the invisible clothes line they seem to have some connexion and continuity." (Kenneth Grahame, 1859–1932)

Declaration

I declare that this thesis has been composed by myself and that the research reported herein is my own. This thesis complies with all the regulations for the degree of PhD at the University of Edinburgh, and falls below the requisite word limit specified.

Bernard Jones

September 1996

Acknowledgements

This work would have been impossible without the help, support and input of many more people than there is space to acknowledge in this section. Bearing this in mind, the people I would like to thank here are Ted Briscoe, for initially suggesting this area of research; my supervisors Lex Holt and Henry Thompson; Shona Douglas, for her high-speed proof-reading and invaluable comments on content and style; my friends and colleagues from the Centre for Cognitive Science and the UK Artificial Intelligence Society (AISB); the Economic and Social Research Council, for funding my studies; and the members of the punctuation community all over the world, in particular the participants of the 1996 ACL/SIGPARSE Workshop on Punctuation in Computational Linguistics.

In addition, I would like to thank my family for putting up with it all — in particular my parents, for all the things they have done for me during my life; and Sharon, without whom this work would all have been much easier, but much less rewarding!

Contents

List of Figures

List of Tables

one Introduction

…then, in 1940, I luckily arrived in the United States, unable to speak a word of English of course, which was a handicap because I found that most people over here do, and I tried to learn the language. I got along as time and a half went by, and I picked up a few words here and there, mostly there because I hadn't been here yet, but I found that people who speak English sometimes do not understand each other too well, due to the fact that they do not use punctuation marks when they talk, and that is why I invented phonetic punctuation, which means that while we talk we will integrate punctuation marks by giving them sounds, so that we can underline our sentences as well when we speak as we do when we read and/or write.
(Victor Borge, Phonetic Punctuation)

Since the 1950's, many computational systems have been designed for the processing and understanding of natural languages, and many theoretical treatments have been developed. The systems have been designed for applications such as automatic translation between languages, extraction of key information from short messages, analysis of language to determine its syntactic structure and/or semantic content, and generating computational output in a naturally understandable linguistic form. Surprisingly, however, very few of these computational systems have made any sensible use of punctuation. Instead, systems designed to analyse text tend simply to strip out all the punctuation marks that occur, and those designed to generate text tend to generate it free of any punctuation marks. The only punctuation mark that seems to be a general exception to this is the full-stop, whose presence in input text can be detected to signal the end of a sentence, and which can be placed at the end of generated sentences.

However, if we look at real instances of natural language, in a newspaper or novel, for example, we are likely to find a great many punctuation marks occurring. It should be made clear at this stage that the current investigation will be focussing almost exclusively on the punctuation system as used in the English language. The fact that real text includes punctuation marks is in itself enough to render the computational approaches that ignore punctuation unsatisfactory, since it will mean that any computer-generated text will look very different from

Let us begin with a confession. I have not watched the first round of **University Challenge** (BBC2) with anything approaching my usual dedication and before we go any further I had better explain why. I cannot answer the questions any more. Over the past few weeks, every time I have dipped in for a quick starter for ten, Jeremy Paxman has asked the sort of question that sends me scurrying over to Des O'Connor on ITV. An almost audible sneer seems to follow: "Too tough for you, huh?" It is. Now this would be a shaming enough admission for any graduate, but it is doubly so for a television critic. For when *University Challenge* returned last year, sans Bamber and avec Jeremy, my how we all scoffed. Easy, peasy, lemon squeezy we all said. They don't make questions like they used to, we all said. And my wasn't it pitiful, we all said, how those eager little undergraduate eyes lit up every time a question on pop music came along? Last night it was my once eager, once undergraduate eyes that waited in vain for a nice question about the Sex Pistols or Spandau Ballet. It never came — the flame of academia, that had once burnt so brightly (well, I always thought so), flickered and went out.	Let us begin with a confession I have not watched the first round of **University Challenge** BBC2 with anything approaching my usual dedication and before we go any further I had better explain why I cannot answer the questions any more Over the past few weeks every time I have dipped in for a quick starter for ten Jeremy Paxman has asked the sort of question that sends me scurrying over to Des OConnor on ITV An almost audible sneer seems to follow Too tough for you huh It is Now this would be a shaming enough admission for any graduate but it is doubly so for a television critic For when *University Challenge* returned last year sans Bamber and avec Jeremy my how we all scoffed Easy peasy lemon squeezy we all said They dont make questions like they used to we all said And my wasnt it pitiful we all said how those eager little undergraduate eyes lit up every time a question on pop music came along Last night it was my once eager once undergraduate eyes that waited in vain for a nice question about the Sex Pistols or Spandau Ballet it never came the flame of academia that had once burnt so brightly well I always thought so flickered and went out

Figure 1.1: Extract from *The Times*, 23/01/96, with and without punctuation

natural text. The function of punctuation marks is more than just orthographic decoration, however, as the marks help readers and writers to understand and structure the text better. Indeed, it seems almost inconceivable to be able to read or produce manually any significant body of meaningful text without the assistance of punctuation marks, since this would either reduce the text to incredibly basic declarative sentences, or render it almost totally incomprehensible. Figure 1.1, for example, contains an extract from a television review in *The Times* newspaper (Bond, 1996), reproduced both with and without its punctuation. The removal of the punctuation makes the second copy of the extract very much more difficult to read than the original — indeed the fact that sections of the unpunctuated version are still comprehensible at all is probably due in the main part to the strict editorial style guidance used in the production of newspapers.

However, the problem of presenting or reading text without punctuation is not just one of

increased difficulty of comprehension. The punctuation present in written language also has a more linguistic role to play, so that individual punctuation marks contribute to the structure and meaning of the text in which they occur. To illustrate this with a fairly basic case, the punctuated sentences in example (1.1), taken from figure 1.1, have a very different meaning to the corresponding unpunctuated text, shown in (1.2).

(1.1) I have not watched the first round [...] and before we go any further I had better explain why. I cannot answer the questions any more.

(1.2) I have not watched the first round [...] and before we go any further I had better explain why I cannot answer the questions any more

Since punctuation does seem to have an important part to play in producing and understanding written language, it seems rather counter-intuitive that so few natural language processing systems have addressed the issue of using punctuation. As mentioned, consideration of the full-stop is fairly wide-spread, but it is rare to find systems that make a more involved use of other punctuation marks. There are, of course, some approaches that have made heavier use of punctuation, for example (Dale, 1990; Hovy and Arens, 1991; Dale, 1991; Douglas and Dale, 1992; Pascual, 1993; Fornell, 1996), but these approaches are by far the minority.

The reason why the use of punctuation in systems is not more widespread, and the reason why those systems that do make use of punctuation tend to do so in an ad-hoc and idiosyncratic manner, is that there has been no proper account of how punctuation works on which a computational treatment could be based. Since the construction of such an account, and hence the computational treatment, is almost invariably a far greater task than the design of the rest of the system, punctuation has tended to be omitted.

The obvious place to turn to for an account of the function and use of punctuation is the field of linguistics. However, as Nunberg (1990) writes:

> With few exceptions, the extensive literature on written language and writing systems has almost nothing to say about punctuation, from either a historical or theoretical point of view. There is little reason to revise the observation that Gleason (1965) made more than twenty years ago, to the effect that "no appreciable amount of research has ever been devoted to [questions of punctuation]... Indeed, there is available very little descriptive data on how the English, or any other, punctuation system is actually used. The large volume of published material which is available is predominantly normative..."

There are several possible reasons for this lack of attention, chief among which are the relatively recent formalisation of punctuation systems within written language, and the general belief that "Punctuation is in large part a system of conventions the function of which is to assist the written language in indicating those elements of speech which cannot be conveniently set down on paper: chiefly pause, pitch and stress" (Markwardt, 1942).

However, more recently there has been an upturn in interest in punctuation, and it is being seen as a system in its own right, performing similar functions for written language as

intonation and prosody perform for spoken language, but being more than merely an orthographic representation of those functions. Linguistic treatments of punctuation have been produced by, among others, Meyer (1987) and most importantly, by Nunberg (1990), from whose monograph *The Linguistics of Punctuation* the above quotation is taken.

Unfortunately for the computational domain, however, neither of these works can really offer a *theory* of punctuation that would be suitable and sufficient as a basis for the development of a computational treatment. Therefore the goals of the work described in this thesis are to produce a theory that will hopefully serve as a suitable starting point for the development of computational treatments of punctuation that can be integrated into Natural Language Processing systems.

1.1 What Constitutes Punctuation?

The content of the unpunctuated text segment in figure 1.1 poses several important questions, chief among which concerns the criteria that are used to determine what symbols and features constitute punctuation, and which are hence used to determine which facets of the orthographic text have been removed. In this case, the criteria used were just to treat all the non-alphanumeric symbols as punctuation, and hence remove all those features. This is based on the dictionary definition of punctuation, such as the following, from (Hanks et al., 1986):

> **punctuation** *n.* **1.** the use of symbols not belonging to the alphabet of a writing system to indicate aspects of intonation and meaning not otherwise conveyed in the written language. **2.** the symbols used for this purpose. **3.** the act or an instance of punctuating.
>
> **punctuation mark** *n.* any of the signs used in punctuation such as a comma or question mark.

However, this is by no means the only view of what constitutes punctuation, and is not even necessarily the view that is generally held. Therefore it becomes obvious that before any discourse about punctuation can be usefully held, it is first necessary to determine a frame of reference regarding the identity of the punctuation which is being talked about.

The most obvious facet of punctuation, and indeed the set of features which are usually referred to when punctuation is discussed, are those that I define as the **inter-lexical** punctuation marks, i.e. those that occur between lexical items (words) in the text . This set of marks, as shown in (1.3), typically includes such features as the colon, comma and full-stop.

(1.3) . , ; : — ( ) " " ? ! ...

(1.4) @ = + # $ & * /

There are other non-lexical items that can occur between words, however, as shown in (1.4), and although symbols of this type are not conventionally regarded as 'punctuation marks' in the most conventionalised sense, there is no reason why they should not be considered as such in the widest sense, since the marks perform similar functions to the more conventional

inter-lexical punctuation. Admittedly, most of these symbols will have a very high semantic and lexical content (in many cases the lexical content is so great that the symbol is read out as a word), but similarities in function between these symbols and the more conventional ones of (1.3) are still clear. Sentences containing emphasis as demonstrated in (1.5) are becoming more and more common in written media, in particular those media that do not allow complex typeface usage, such as electronic mail.

(1.5) Are you sure you ∗really∗ mean that?

In addition to the inter-lexical punctuation marks, there are other graphical symbols present in text that are conventionally seen as being part of 'punctuation'. These are the marks that occur within words and include the hyphen, the apostrophe and the abbreviation-marking point (1.6).

(1.6) music-hall it's T.S.Eliot

(1.7) sub- and inter-lexical parents' concerns I.B.M.

These marks, which I define as **sub-lexical** (which, confusingly, can sometimes appear in inter-lexical positions (1.7)) tend to have functions whose meaning, use and mechanism are relatively well understood. Hyphens, for example, are conventionally used in some compound words (1.8a), in certain adjectival phrases (1.8b), and to signal word-breaking at the end (and, in certain linguistic conventions, the beginning) of a line (1.8c). Apostophes can be used to mark elided material within an orthographic word (1.9a) and to signal the genitive (possessive) case (1.9b).

(1.8) a. mother-in-law labour-saving X-ray

b. It was a very run-of-the-mill meeting.

c. Traditionally, the compositors and type-setters were trained in the division of words...

(1.9) a. he's haven't won't've they'll

b. ladies' golf Jane's toy the children's toys

It is worth noting that, although the mechanisms under which sub-lexical marks operate tend to be well-understood, this still does not mean either that all such mechanisms have been described (c.f. (1.10) and (1.11)) or that they are all used in the conventionally correct manner (c.f. "the visible and often highly-ornate manifestations of signwriters' confusion up and down the land" in relation to apostrophes (Jarvie, 1992)). In addition, even when used according to convention, the sub-lexical marks can give rise to confusion and ambiguity, as when for example the apostrophe or abbreviative point occur in word-final positions, and are mistaken for or interact with the graphically similar marks of inter-lexical punctuation, such as the final single-quote and full-stop. The fact that such relatively well-understood marks of punctuation can give rise to such problems only serves to underline the necessity for studying the other, more complicated punctuation marks to determine their function and use.

(1.10) "I'll b-b-bring out the ca-ca-casserole."

(1.11) "Please speak c-l-e-a-r-l-y into the microphone."

Our definition of punctuation so far, then, is that it includes all non-alphanumeric visible characters. However, we can and should take this definition further. A full-stop, which is naturally regarded as being part of punctuation, serves to mark the end of a sentence. Since a capital letter marks the beginning of a sentence in almost exactly the same fashion, there is not really any argument for excluding the capitalisation function from the set of punctuation phenomena. Non-sentential uses of capital letters, for proper nouns, etc., are analogous to other sub-lexical punctuation phrenomena, such as the use of the dot in marking abbreviations. That sentence-initial capitalisation can be counted as a punctuation phenomenon is an important realisation, since it means that our notion of punctuation is diverging from the generally accepted definition, of a set of concrete graphical symbols. The fact that a graphical procedure has been added to the set subtly changes the nature of punctuation from being just a set of characters to also including actions that affect the graphical manifestation of the surrounding text.

From a computational point of view, then, this means that punctuation changes from a well-defined set of ASCII characters to anything that is not an alphanumeric character (and since we are regarding capitalisation as a punctuation phenomenon, this definition of alphanumeric should exclude upper-case letters). Therefore all the other orthographic processes that are present in text should also be seen as punctuation phenomena. I define these these processes as **super-lexical** punctuation, examples of which are paragraphing, underlining, italicising and superscripting, amongst other structural processes. To illustrate that these super-lexical operations have a similarity to conventional, inter-lexical punctuation, consider the sentences in (1.12), based on some from Nunberg (1990). Although a little hard on the eyes, since we are not used to seeing information presented in this way, sentence (1.12b) can still be read and analysed correctly (the line-break can be a little confusing, however).

(1.12) a. He assigned Silas Marner — a book that Jean, who spoke little English (and read less), could barely understand. We read it anyway.

b. He assigned *Silas Marner* *a book that Jean* *who spoke little English* **and read less** *could barely understand*. We read it anyway.

Our working definition of punctuation, then, is any facet of the text that is not a lexical item or number, and using this definition causes the newspaper text from figure 1.1 to be represented as in figure 1.2. Now lacking even capitalisation, typeface and paragraph phenomena, this unpunctuated text is very difficult to read and understand. However, there is one additional non-lexical facet of text that we are ignoring in using this definition, and that is the inter-word space. This is undeniably a significant feature, since it allows us easily to recognise what constitutes a lexical item — which could otherwise prove problematical in certain circumstances (1.13).

let us begin with a confession i have not watched the first round of university challenge bbc2 with anything approaching my usual dedication and before we go any further i had better explain why i cannot answer the questions any more over the past few weeks every time i have dipped in for a quick starter for ten jeremy paxman has asked the sort of question that sends me scurrying over to des oconnor on itv an almost audible sneer seems to follow too tough for you huh it is now this would be a shaming enough admission for any graduate but it is doubly so for a television critic for when university challenge returned last year sans bamber and avec jeremy my how we all scoffed easy peasy lemon squeezy we all said they dont make questions like they used to we all said and my wasnt it pitiful we all said how those eager little undergraduate eyes lit up every time a question on pop music came along last night it my once eager once undergraduate eyes that waited in vain for a nice question about the sex pistols or spandau ballet it never came the flame of academia that had once burnt so brightly well i always thought so flickered and went out

Figure 1.2: Extract from *The Times*, 23/01/96, without any punctuation phenomena

(1.13) a. **stab led** vs. **stabled**

b. **friend ship** vs. **friendship**

c. **mans laughter** vs. **manslaughter**

d. **Her man** vs. **Herman**

Indeed, when the historical development of punctuation (discussed in greater detail in chapter 2) is examined, early producers of written language such as the early Greeks and Romans did not bother to separate words with a space, or otherwise indicate the boundary between one word and the next (Parkes, 1992; Casson, 1988). Indeed, for a considerable time, no punctuation was used in texts at all. Casson writes:

> Count your blessings, readers of today. If you were living during the years of Greece's glory or Rome's grandeur, that last sentence would have looked like this:
>
> COUNTYOURBLESSINGSREADERSOFTODAY
>
> And, if it had been an inscription on a public monument in Athens, say, where the watchword was balance and proportion above all, it would have looked like this:
>
> C O U N T Y O U
> R B L E S S I N
> G S R E A D E R
> S O F T O D A Y[1]

Therefore, if we were to be entirely rigorous, the inter-word (and inter-sentence) spaces should really be considered as much part of punctuation as a paragraph-break or a hyphen. To

[1]Except, of course, they would not have used roman letters or U!

letusbeginwithaconfessionihavenotwatchedthefirstroundofuniversitychallengebbc2withanythingapproachingmyusualdedicationandbeforewegoanyfurtherihadbetterexplainwhyicannotanswerthequestionsanymoreoverthepastfewweekseverytimeihavedippedinforaquickstarterfortenjeremypaxmanhasaskedthesortofquestionthatsendsdsmscurryingovertodesoconnoronitvanalmostaudiblesneerseemstofollowtootoughforyouhuhitisnowthiswouldbeashamingenoughadmissionforanygraduatebutitisdoublysoforatelevisioncriticforwhenuniversitychallengeretur nedlastyearsansbamberandavecjeremymyhowweallscoffedeasypeasylemonsqueezyweallsaidtheydontmakequestionsliketheyusedtoweallsaidandmywasntitpitifulweallsaidhowthoseeagerlittleundergraduateeyeslitupeverytimeaquestiononpopmusiccamealonglastnightitwasmyonceeageroneundergraduateeyesthatwaitedinvainforanicequestionaboutthesexpistolsrrspandauballetitnevercametheflameofacademiathathadonceburntsobrightlywellialwaysthoughtsoflickeredandwentout

Figure 1.3: Extract from *The Times*, 23/01/96, without punctuation or spacing

give some indication of what the lack of this particular punctuation mark would do to comprehensibility, figure 1.3 reproduces the text from figure 1.2 totally devoid of even the inter-word space punctuation mark. The spacing feature is such an obvious and essential one, though, that it can almost be passed over in most accounts of punctuation since its presence is taken for granted. Hence, what we consider to be part of punctuation for the purposes of this account is the other non-lexical features of the text.

We are dealing, then, with the features of sub-, inter- and super-lexical punctuation. However, sub-lexical punctuation, as already discussed, tends to operate in reasonably well-defined and well-documented ways and since its location is within words it is more suitably dealt with in morphological systems or the lexicon. Of course, being well-defined and documented does not mean that this is an uninteresting or unrewarding area of research. Indeed, many challenging problems manifest themselves, especially regarding the interaction and confusion between sub-lexical and inter-lexical features, for example as discussed in (Palmer and Hearst, 1994). However, the function of these marks is separate from the function of the rest of punctuation, in so far as the sub-lexical marks change the meaning of the words that contain them, wheras the inter-lexical and super-lexical marks change the manner in which the words combine to produce an overall meaning or purpose. Therefore sub-lexical marks will not be focused on, except peripherally, in this investigation.

The problem with the more structural super-lexical marks is one of representation. It would be necessary to employ a standardised text mark-up language, such as SGML or LaTeX code to properly represent these, and unfortunately no such standard representation is currently in common and generalised usage. Hence consideration of these marks would be technically difficult, and so they too will not be considered in the current investigation. There is research being done in the area of super-lexical punctuation (Douglas and Hurst, 1996; Pascual and Virbel, 1996), but this tends to cover only very specific phenomena (e.g. tabulated information, chapter headings). In the absence of a sufficiently standardised mark-up convention, consideration of the full range of super-lexical phenomena would be too difficult a task. However, as the functions of certain super-lexical phenomena (paragraphing, sectioning, etc.) are similar

to, or at least comparable with, the function of some inter-lexical marks, as illustrated in (1.14), the results of investigating inter-lexical marks of punctuation should without too much trouble be easily adaptable to also cover some super-lexical punctuation phenomena at a future stage.

(1.14) a. Jack ran to his house. He was late.
John was worried. ¶ Jim is working.

b. Fred (my best friend) is coming over tonight.
James *a man I trust* will call you.

1.2 Choosing an Approach to Punctuation

Following the considerations of the previous section, this account will concentrate on the inter-lexical range of punctuation marks and phenomena. This therefore will not only tackle those marks which are most commonly thought of under the term 'punctuation', but is also likely to yield results that should be easily applicable to a wide range of computational applications.

The issue that should be addressed first, however, is how the field of punctuation will be approached. Conventional studies and investigations to date — mainly in a purely linguistic or pedagogic/normative field — have either tackled the issue of punctuation in a mandatory and expected place (the full-stop, for example), or have described a set of rules and guidelines for the possible, allowable locations of punctuation and described the use and possible function of marks occurring in those locations.

The reality of the situation is a little more complex, however. Punctuation is almost unique amongst linguistic phenomena in that it is an area where there is a huge amount of argument, discrepancy and uncertainty in usage between people. Trask and Wade (1993) begin their guide to punctuation usage with:

> Why should you learn to punctuate properly? After all, many people have made successful careers without ever learning the difference between a colon and a semi-colon. Perhaps you consider punctuation to be an inconsequential bit of decoration, not worth spending your valuable time on. Or perhaps you even regard punctuation as a deeply personal matter — a mode of self-expression not unlike your taste in clothes or music.

Teaching people the fundamentals of punctuation is a laudable aim, but very frequently one that either fails altogether or is imperfectly achieved. Usage of punctuation, even among supposedly linguistically-astute people, is very idiosyncratic. This, of course, is the reason that editors' style guides constrain the usage of punctuation in publications — to try to maintain a constant usage over time and different authors. Usage of these guides, again, is laudable since it achieves written output that is not cluttered with unnecessary punctuation, and yet is clear and understandable. One thing that is instantly obvious, for example, from reading the newspaper text in figure 1.1 is the relative paucity of the punctuation contained in it. There is only one colon, no semi-colons and a reasonably small number of commas — far less punctuation than we would expect to see in a piece of un-edited text produced by a reasonably competent writer.

The idiosyncrasy of punctuation usage, however, is a problem when it comes to analytical systems. There is little to be gained from blindly applying the normative and prescriptive approaches to the analysis of text, since we are trying to extract as much meaning as possible from the input. Our initial assumption must be that the text to be analysed is coherent and has content — hence for an analysis to fail because some idiosyncratic punctuation is encountered would be totally counter-productive. Instead, we have to realise that the usage of punctuation marks can be divided into three categories. There are those situations where usage of punctuation is mandatory, for example at the end of a sentence, and those situations where usage of punctuation is impossible (1.15). But in addition to these two, there is a third set of situations where punctuation is possible, but not necessary. In these situations, valuable information can be extracted from the presence and variety of a punctuation mark, but if the mark is absent the analysis can go ahead anyway and the best can be made of the situation.

(1.15) a. The biggest, tree in the forest is over here.

b. My mother is, going to kill me.

c. I wrote, an essay yesterday.

This is how any workable analysis and computational implementation of punctuation function should operate: not necessarily to expect the presence of punctuation, but to be able to extract the maximum amount of useful information from the punctuation that is present in whatever text is to be analysed. Of course, the implicit assumption made here is that none of the text will have any punctuation in the 'impossible' positions, but since punctuation would only ever be placed there by totally incompetent writers, this should not prove too problematic. The alternative approach, to avoid even this problem, would be to ignore any punctuation placed in the 'illegal' positions, but that then introduces a layer of prescription regarding what constitutes a legal and an illegal position, and it seems desirable to avoid as much prescription as possible.

Bearing this in mind, then, what this research is trying to achieve is an account of the variety, use, form and function of inter-lexical punctuation marks in text, and the synthesis of a theory to describe this that will be applicable in a computational domain.

1.3 The Structure of the Thesis

This thesis will begin with survey of both the historical development of punctuation marks and their uses, and a review of literature on modern approaches to punctuation. It is necessary to have an insight into the development of punctuation since so much of the variety and functional characteristics of punctuation marks trace their origins to historical practices. The historical development also shows the operation and reasons for the creation of the whole punctuation system, in addition to giving usage examples for the individual marks. The current literature builds on this developmental data to show how punctuation marks, and the whole punctuation system, are viewed and used in contemporary English. This will then provide a platform upon which the rest of the investigation can be based.

Even if a theory of punctuation can be synthesised, this is no guarantee that such a theory would be computationally implementable, or tractable. Therefore the next chapter of the thesis discusses work that justifies the study of punctuation by showing that it can yield considerable benefits to computational language processing systems.

Following the results that accounts of punctuation can be of use to computational systems, the path is clear for an investigation of the punctuation system and the development of a theory. The investigation is split into three facets: an exploration of the actual punctuation marks themselves (what they are, their frequency of occurrence, how they combine with one another), an exploration of the syntactic function of punctuation in text (from a variety of approaches), and an account of the semantic functionality of punctuation marks.

This data is then combined in a chapter that constructs a taxonomy of punctuation marks, their variety and their function, which is followed by a synthesis of all the information obtained into a theory of punctuation that provides a grounding from which computational treatments can be developed.

Before summarising the research, the possibilities for transferring the applicability of the theory to languages other than English are also explored.

two The History of Punctuation

'Mr Speaker, I said the honourable member was a liar it is true and I am sorry for it. The honourable member may place the punctuation where he pleases.'
(R B Sheridan, in Parliament)

We have long passed the Victorian era, when asterisks were followed after a certain interval by a baby. (W Somerset Maugham, 1874–1965)

Having obtained a working definition of what constitutes punctuation in present-day orthography, it is important to examine the history of punctuation to gain insights into the variety, composition and use of modern punctuation marks. As a facet of orthography, punctuation is a relative newcomer, only having appeared since around 200 BC (Parkes, 1992). Graphical writing, on the other hand, traces its origins back as far as the Upper Paleolithic era, from 35,000 to 15,000 BC (Diringer, 1962; Gelb, 1963).

Of course, the writing of this period was not what we would associate with the term in the present day; rather it was *pre-writing* of a sort that showed the need of its producers to communicate in some sense with their fellows, but not yet showing those intentional facets of *conscious writing* as a system of human intercommunication by means of conventional visible marks (Senner, 1989a). This pre-writing consisted of such phenomena as rock paintings and carvings on rock and bone, including concrete imagery (e.g. pictures of a hunt, or a rain ceremony), abstract work (e.g. geometric and regular or decorative patterns) and symbolic images (abstract symbolisations of concrete entities, e.g. a circle with eight rays to symbolise the sun).

However, even graphical systems that we recognise as writing predate punctuation by a large margin. Plain and complex tokens were developed and used in the middle-east as counters for accounting purposes in around 8000 and 3400 BC respectively (Schmandt-Besserat, 1986). Plain tokens consisted just of small 3-dimensional geometric counters, made

of clay, in shapes such as spheres, discs, cones, tetrahedra, etc. Complex tokens consisted of similar shapes (and also more complicated ones, such as ovoids, bent coils, triangles, representations of tools and animals, and small vessels) but these additionally bore markings on their surfaces. These markings consisted typically of linear pattern, notches and dots that were traced or impressed with a stylus, or in rare examples were represented in raised clay. These tokens, both complex and simple, were used not only to denote numbers and quantities, but also to indicate what was being measured (i.e. grain, sheep, etc) (Schmandt-Besserat, 1989). Importantly, these tokens came to exist not just as physical entities that could, for example, be threaded onto a string or kept in a box as a permanent representative record, but could also be used to mark soft-clay tablets, leaving a 'negative' image of the token behind to signify the same as the token itself would signify.

This led directly, by approximately 3200 BC, to the development of the cuneiform script used in Mesopotamia. Based initially on the imprints of the plain tokens in clay tablets, and wedge-shaped impressions made with a stylus to represent the complex tokens, it was possible to broaden the vocabulary of cuneiform script from its base of starkly utilitarian accounting-based symbolism (Green, 1989). The script became more powerful and flexible in transcribing messages, and therefore its scope branched out into narrative and creative uses. Poetry, prose, chronicles, dictionaries, religious and scientific works were all produced as this pictographic writing form evolved into a regular, nonpictorial script. As it developed, this writing also acquired a phonetic element, with the signs possessing multiple values, logographic and syllabic.

By 1800 BC, a number of logographic/phonetic scripts were in use in the Near East — Sumerian, Akkadian, the Egyptian Hieroglyphs, etc. — but these were all extremely complicated. Typically, these systems of writing would consist of several hundred signs, making it a formidable task to learn and use them, and in effect making writing the monopoly of royal and religious courts (Cross, 1989). This provided an appropriate background for the development around 1800 BC of the first alphabet: the Old Canaanite. This was a system of extreme simplicity, consisting of only 27 or 28 symbols, with each symbol representing a consonontal phoneme (vocalic phonemes did not exist in the early alphabet), and indeed this was the only alphabet ever invented. All subsequent alphabetic writing systems, e.g. Latin, Greek, Arabic, Hebrew, were developed from this original Old Canaanite system (Naveh, 1982).

By 1100 BC the Arameans had devised a rudimentary system for representing certain vowels, but the first full system of signs for vowels did not come until developed by the Greeks between 900 and 800 BC. The Latin writing system was developed from the Greek system in around 700 BC, and it is the Latin writing system in which the use of rudimentary punctuation can first be observed.

The culture of the ancient world was dominated by the ideal of the eloquent orator. The ability to make a good speech in public was considered essential. This encouraged the emphasis on a verbal response to the written word: texts were almost invariably read out aloud either in a murmur if the reader was alone, or, more ideally, as expressive declamation in public (Parkes, 1992). However, authors (or rather their amanuenses and scribes) rarely if ever inserted any symbols or spacing into the text other than the letters comprising the words

contained therein. Thus text consisted of a long string of letters, with no spacing even between words, called *scriptio continua* in Latin.

This text therefore appeared on the page with far less information in it than would have been required to let a reader read it out at sight. Ambiguities and mistaken word-division would have led to mistakes in the interpretation of the writing, e.g. (2.1). By 200 BC, in the Roman republic, the first rudimentary instances of punctuation use can be observed, as layout features were employed by scribes to indicate major divisions or sections of a text (Müller, 1964). Not long after, by 100 or 50 BC the first letters of new paragraphs were set out to left of the line and enlarged (*litterae notabiliores*, which were the precursors of contemporary capital letters). By 50 AD spaces were used by some scribes to mark major pauses within paragraphs, and by 100 AD Latin manuscripts had the words separated by interpuncts (2.2), a practice the Romans had derived from the Etruscans.

(2.1) *conspicit ursus* (a bear espies)
conspicitur sus (a sow is espied)

(2.2) SICVLORVM·VRBES·SIGNIS·MONVMENTISQVE

However, after the end of the first century, the use of interpuncts in Latin texts died out and scribes imitated the Greek practice of writing without separating words or indicating any pauses within a major section of text. Writing thus reverted back to *scriptio continua*. The merit of this form of writing, to the ancient Romans, was that it presented any reader with a neutral text. It was felt that introducing graded pauses into the text involved an act of interpreting the text which should more properly be reserved for the reader. Trying to read a text written in *scriptio continua* therefore involved careful preparation, and identification of the elements of the text: letters, syllables, words, phrases, etc.

The ability to perform this analysis of continuous script was not an easy one, and therefore was carefully taught during secondary education. The principle teaching method was via the *praelectio* ("an exposition of the text designed to make pupils follow the written word with ease and accuracy" (Parkes, 1992)). This involved either the teacher or the pupil preparing the text by inserting marks to link and separate words and to indicate position and duration of pauses. These marks were the first true symbols of punctuation, and served both a disambiguating and a prosodic function.

Scribes would enter some of the major marks, such as the ℵ that marked the begining of a new head in the argument (*kaput*), whilst readers would insert the minor ones such as oblique strokes / to indicate minor pauses or the *simplex ductus* ⁊ to separate some preceding matter from following matter which had been run together although unconnected (e.g. to separate items in a list, such as place names or events).

Thus punctuation was, to a great extent, left to the reader to insert as he required it or thought it proper. By 400 AD, the concept of punctuation marks began to change from an ancillary device to help inexperienced readers to a legitimate addition to the text to ensure that it was properly understood. Favourite texts were annotated by scholars for the benefit of others, so that they could easily and correctly interpret them. At the same time, therefore, standardised systems of punctuation were developed so that the marking up of texts could

be unambiguously understood. One such system was that of Donatus (Holtz, 1981), who divided sentences using three similar marks, at different heights to indicate an ascending level of importance. A low point . indicated a minor medial pause, a mid-level point · indicated a major medial pause and a high point · indicated the final pause after the sentence was complete. Interestingly, the Latin terminology for the phrases preceding these marks in each case is *comma*, *colon* and *periodus*. Thus the names used for modern punctuation marks are derived from the Latin terms for the types of phrasal-entities their ancient Latin equivalents used to come after.

One of the first Latin books to be written with punctuation included for its readers (as opposed to those that were marked up after they had been written) was the version of the Bible produced by St. Jerome in around 400 AD. The reasons for its preparation in this manner were Jerome's dissatisfaction with the need for so much correction in earlier manuscripts, and the need for accurate, invariant interpretations of the book that was of paramount importance to the Christian religion (Parkes, 1992). Additionally, a major part of the reason that the book was producable with punctuation was that Jerome could be sure that his scribes would transcribe his punctuation faithfully, something that could not always be relied upon at that time. Jerome punctuated his Bible by, in addition to the use of *litterae notabiliores*, changing the layout of the text. It was structured *per cola et commata*, or the laying out each consituent part of the sentences on new lines, as illustrated below in a translation of the Vulgate Bible (in the original, the longer lines would have been split and the wrapped lines indented from the margin).

> Blessed is the man who hath not walked in the counsel of the ungodly
> and hath not stood in the way of sinners
> and hath not sat in the seat of the scornful
> But his will is the law of the Lord
> and in His law shall he meditate day and night
> And he shall be like a tree transplanted close by the streams of water
> that will bring forth its fruit in due season

This innovation established a precedent, and because of Jerome's authority this new way of writing became increasingly popular. By 600 AD, however, the ideal of the orator no longer dominated literary culture, and in addition forms of spoken language began to diverge increasingly from that represented by the Latin texts. Therefore it was more important than ever to clearly indicate the meaning of texts in what was becoming, to many readers, a foreign language. At the same time, the existence of writing as a linguistic medium in its own right (rather than just a transcription of the spoken word) was being recognised and silent reading began to increase in popularity. This change of attitude was first witnessed by Isidore of Seville, who documented and discussed the changing status of the written word and wrote one of the first accounts of punctuation, *"De posituris"*, based on the system of Donatus, but also describing many other symbols in regular use (Parkes, 1992; Fontaine, 1959).

Isidore's writings were circulated widely throughout Europe, and influenced writing and punctuation at least until 900 AD. From 700 AD onwards, use of the ancient book scripts began to decline and new scripts, based on cursive and calligraphic writing styles, began to be used.

Furthermore, other languages besides Latin began to be written down. The scribes of Ireland were particularly influential in the development both of writing and punctuation. They not only invented new symbols for abbreviations, but also subdivided Latin sentences better and more clearly by separating words with spaces and deriving new punctuation marks that were also better suited to the new scripts than the marks from the ancient ones. The Irish also began the practice of grouping marks of punctuation together to signify new meanings (e.g. three comma-shaped marks marking the end of a section, and double or single marks marking subdivisions inside the section) (Lowe, 1971).

The Irish manuscripts also show the conceptual linking of punctuation and decoration within the text, as the scribes saw these as two aspects of the same thing: the presentation of a text which facilitates its use. This led to the development of elaborate, decorated *litterae notabiliores* (capital letters), 'carpet pages' that contained nothing but elaborate decorations, and the use of different colours in writing (certain punctuation marks were rendered in colour for decorative and distinguishing reasons). These scribal practices, placing more emphasis on the visual impact of punctuation and layout, and therefore producing greater clarity, made a significant contribution to the development of punctuation.

The Anglo-Saxon scribes learned these new practices from the Irish and sought to improve them further. Legibility and clarity were still rather poor in certain Irish manuscripts owing to their inclination to compress as much information as possible onto a single page (Brown, 1982). The Anglo-Saxons strove to correct these faults by restricting the number of variant forms of letters and producing new, more-standardised and legible calligraphic letter forms. Word-separation was also introduced into copies of *scriptio continua* works, to improve clarity and comprehension. A more sophisticated notion of page design was also developed, partly by restricting the decorative innovations of the Irish, but also by the development of some particular, more rigid conventions for producing texts.

The use of different scripts to distinguish passages of the text was also introduced by the Anglo-Saxons; thus two different script styles might be used in a commentary to distinguish the original text from the passages discussing it. The study of ancient texts also, ironically, caused some Anglo-Saxon scribes to reintroduce ancient marks of punctuation to their own texts, such as the ivy-leaf (or *hedera*)  to separate two sections or distinguish normal text from commentary, the *diple* > to indicate quotations, and ☙ to mark a new chapter. The majority of scribes, however, simply adapted the Irish system of multiple punctuation marks indicating the significance of the pause (Parkes, 1992).

These new practices, developed by the Irish and Anglo-Saxons by 800 AD (*insular scripts*), were then re-exported via scribes and books back to the continent. In the 780's one Anglo-Saxon scholar, Alcuin, began to assist Charlemagne to promote liturgical, educational and administrative reforms throughout the Holy Roman Empire (Bischoff, 1981a; Bullough, 1973). These reforms included the revival of ancient Latin for use in the business of Church and State as distinct from the *lingua romana* that was used in everyday life. The standards required for such a return to ancient 'latinity' required wide reading and therefore large quantities of copies of ancient books. Therefore there was a certain amount of confusion between the ancient scripts and conventions and the newer insular ones. Scribes began to insert punctuation marks on

their own initiative, using both the insular punctuation marks and the more ancient ones. However, at the same time, further reforms were carried out to simplify letter forms to *litterae absolutae* — invariable letters — so that there was only one form of each letter, and that form included the minimum number of distinctive graphic features to enable differentiation from other letter forms. Thus while the letters of the text had become standardised, the punctuation had not yet achieved this, as the marks had not yet acquired sufficiently clear identities in relation to each other or to the scripts in which they were used. The scribes of Charlemagne had succeeded in assembling a variety of punctuation marks from different sources, and therefore increased the number of possible variants, resulting in the potential for hyper-characterisation. The number of forms needed to be reduced, and their graphic properties and status relative to one-another refined and defined to ease their interpretation and the establishment of their identity (Parkes, 1992).

At around the same time (800 AD) a new system of punctuation was developed, for use in liturgical texts, where the need for adequate, clear and unambiguous punctuation was paramount. This system of symbols (*positurae*) consisted of four marks — to indicate the end of a declarative sentence (*punctus versus* ;), the end of a questioning sentence (*punctus interrogativus* ⁀ or ⁀), a major medial pause (*punctus elevatus* ⁝) and a minor medial pause (*punctus flexus* ⁘) (Bohn, 1887). Apart from liturgical uses, however, the *positurae* did not really catch on until around 1100 AD. The volume of material using the older systems of punctuation was just too large to enable the newer marks to take over, since the punctuation system was usually transmitted on copying along with the text. The *positurae* became more and more popular, however, especially those symbols that did not have a parallel under the old system (such as the question mark). Their main advantage lay in their clarity and ease of recognition, as well as their easy integration with the newer, cursive and calligraphic scripts.

From 1100 AD onwards, with the basic conventions of orthography well-established, scribes began to abandon earlier systems of punctuation in texts in favour of a newer, more standard repertory. This was developed from an amalgamation of the elements of previous systems, and therefore consisted of marks drawn or adapted from these earlier systems. However, there was still no general standardisation of the particular marks used, so that whilst the position in which a mark occurred would be a standard one, the identity of that mark would not necessarily be standard.

Handwriting became more compressed, and so the space available for punctuation marks decreased, favouring the simple marks. Therefore the most common mark of punctuation became the *punctus*, or point. It was used to indicate all kinds of pauses, to separate, introduce quotations and mark abbreviations and titles. Since this led to a great deal of ambiguity, the mark was generally combined with other symbols to distinguish between the various uses, such as with other points, comma-like marks, etc. Combined with these symbols, the height of the point ceased to be important, and thus it became just the graphical shape of the punctuation mark that determined meaning, rather than its vertical position (Hector, 1966).

Punctuation continued to be standardised, both in use and function until the advent between 1400 and 1500 AD of the humanists. Wanting their texts to be read widely and easily, punctuation was of great importance to the humanists, but since they regarded eloquence as a

generative process which fashioned their own expression, each writer had to attend to his own punctuation rather than merely following a predetermined method or model. They tried to bring out both the logical relationships between the subordinate constituents of their sentences and the periodic structure of their discourse (Ricci, 1943), drawing on the widest possible range of symbols available. Thus, as well as the re-reintroduction of some of the ancient Roman symbols, new and varied symbols were also employed to indicate pauses within the sentence, indicating such new functions as parenthesis (indicated by / or :).

The new symbols used by the humanists began to pass into the general repertory of punctuation, and so they became responsible for the introduction of our modern colon, parenthesis, semi-colon and exclamation marks. However, the contribution of the humanists to modern punctuation is not just these new marks of punctuation, but also their attitudes to their usage. They demanded an exact disambiguation of the meaning and function of each punctuation mark used, which led to a stricter system of usage of punctuation marks (i.e. particular marks could only be used in particular, constrained circumstances to achieve a particular effect) that exerted a powerful influence on later writers (Parkes, 1992).

Simultaneously, the punctuation system was undergoing another major change, due to most radical change in the history of orthography: the invention in the 1430's of printing presses and movable metal types. Simple marks were preferred as these were easier to set into metal type and print, and standardisation occurred with the desirability for the reuse of a small number of metal punctuation-mark types. As particular type foundries and type faces gained ascendance, the degree of standardisation increased. By 1600, marks in common circulation included point, comma, colon, question mark, exclamation mark, semi-colon and parentheses. Other marks of punctuation were used, as required by the work in question, and even some novel marks appeared, such as Henry Denham's back-to-front question mark ⸮ to signal a rhetorical question (Tannenbaum, 1931).

The shape and layout of text was also constrained by the use of movable type. A line of text needed to have restrictions as to the vertical displacement of symbols it contained, thus fixing the position of punctuation marks as aligned with text. New conventions also had to be developed to perform tasks such as alignment or reference in commentaries, and new punctuation marks were therefore developed to this end. The apostrophe was used, initially to mark the elision of a vowel, and was retained later even when the vowel no longer appeared in the spoken form of the word. A sign of suspension was also developed, to indicate a point where a speech in a dramatic text had been interrupted. The mark used for this was a series of dashes or points, which has developed both into our modern dash interpolation (suspension of current argument or discourse to begin a digression) and the modern ellipsis, which can now both be used outside the context of speech. Certain other old marks also survived (those that could easily be reproduced in metal type), such as ∗ and † and these were employed for specialised uses, such as marking deliberately omitted letters (e.g. "om∗ssion") and referencing footnotes. As the number of such notes increased, however, it became more standard practice to use letters or numbers to reference them.

By 1700 the notion of the *orthographic sentence* was firmly embedded in writing practice (Levinson, 1986), i.e. the graphical notion of the sentence as we have it now: containing a single

line of argument, and delimited with capitalisation and final-stopping punctuation features. Previously, the situation had been far less clear, with whole paragraphs sometimes being punctuated only with marks less significant than the full-stop (or the full-stop occurring in the middle of what we would consider a sentence), since the punctuation related just to pauses to be made in (silent or aloud) reading.

Additionally, at the same time, quotation marks began to be used to mark quoted speech, rather than the *diple* marks that had been used since antiquity (these had been used by setting one in the left margin alongside each line that contained quoted material). However, there were still many different methods used for denoting quoted material. Richardson, for example (1748), preferred the use of a single in-line diple before quoted material from a letter, and a line-break or dash before direct speech, although he did also use quotation marks. Subsequently, the practice of using a diple or dash caught on in certain punctuation traditions, and is probably responsible for the usage in French of the *guillemets* (≫ ≪) as quotation marks.

The invention of movable type, and its regularisation, has had the effect of freezing punctuation. In the last three centuries, no new marks of punctuation have been developed (although we have made some new combinations, such as "**:-**"), and the use of existing punctuation has to an extent been standardised. Development in punctuation in recent years has rather been to decrease the number of situations in which punctuation is used, to eradicate marks that are not strictly necessary (such as commas at the ends of lines in addresses, and certain abbreviation-marking points) and to specify the function and meaning of punctuation marks more exactly.

2.1 Summary

Thus we have seen that punctuation is a relatively new phenomenon in the history of writing, but that it occupies a crucial place in the interpretation and disambiguation of text. The history of punctuation marks is a curiously cyclical one, however.

Since the relatively free and varied use of punctuation marks by early Roman orators to mark up their personal copies of texts, there has always been a desire to regularise, codify and simplify the prevalent system of punctuation. However, until the advent of movable type and printing, this process of regularisation would be interrupted periodically by the rediscovery of ancient or different punctuation systems that were regarded as interesting and richer than the prevailing system, and the other marks would be incorporated into the prevailing system, rendering it once more complex and potentially ambiguous.

It is possible that this cycle is repeating itself once more in the present day, but in a different manifestation. Due to the restrictions of standardised typefaces and computational character sets, and also due to widespread ignorance of ancient texts and punctuation, the older texts are no longer the sources of 'novel' punctuation. Rather it is the conventionally under-used non-alphanumerical symbols that are present in standard character sets that are now arousing people's interests and finding increasing use in, for example, electronic communications such as USENET articles.

Such development of standard modern punctuation is not at all widespread, however, and therefore it could reasonably be said that in terms of graphical variety, modern punctuation

is static. What development there is tends to be in terms of usage (paring down punctuation usage to only those situations where the marks are absolutely necessary) and meaning (greater specification of the significance of the individual punctuation marks).

Having seen how modern punctuation marks have arisen historically, it is now necessary to examine the role they are considered to play in modern language, in order to be able to incorporate this information in the computational treatment that punctuation needs to receive.

three Approaches to Punctuation

"I am more or less happy when being praised, not very uncomfortable when being abused, but I have moments of uneasiness when being explained."
(Balfour, cited in (Bennett, 1994))

After exploring the historical development of punctuation over the last 2000 years, it is imperative for the development of a theory of punctuation to explore modern approaches to the subject, so that the theory can be based on the 'state of the art'. Additionally, these accounts of punctuation will serve not only to present current literary thinking and principles on the subject, but also to illustrate the state of development that punctuation is currently in (since as discussed in the previous chapter the physical/graphical manifestation of punctuation has changed little in the years since the common usage of the printing press and movable type, so that any development in the system of punctuation now takes place almost exclusively in terms of usage and function).

The need for the development of a theory of punctuation was mandated in the introduction on the fact that there was no coherent, logical treatment of the subject available that gave a good account of the usage and function of punctuation marks and was also suitable as the basis of a computational account. However, this is not to say that there are no accounts of punctuation in existence; there are reasonably large numbers of such works, but although they will serve to provide a useful and valuable background for a theory, these accounts on their own are not sufficient for the formulation of such a theory.

In his recent book review, Humphreys (1993) suggests that accounts of punctuation tend to fall into three categories.

> "There are three sorts of book on punctuation. The first, prototypically authored by retired schoolmasters, is selflessly dedicated to the task of bringing Punctuation to the Peasantry. Somewhat 'hobby-horsical' in tone, they usually contain lengthy

> pleas for the better treatment of the semi-colon, a stop which the author invariably considers to have been 'wantonly neglected' of late. The second sort is the Style Guide, written by editors and printers for the private pleasure of fellow professionals and, in consequence, rarely seen in the open. The most important function of the guides is to state which punctuation convention is to be adopted in those places where the punctuation system as a whole leaves open the possibility of alternatives (e.g. the appropriate handling of full-stops within embedded quotes).
>
> The third, on the linguistics of the punctuation system, is much the rarest of all ... Presumably this paucity reflects a widespread belief that, mere stipulation aside, there really is nothing much to be said of the topic. I would be lying if I said that [the researchers who have produced such works] have stumbled upon the shores of a linguistic America, returning laden with treasure and tales of vast new territories awaiting the attentions of the intrepid researcher; for it is not so. But there are certainly some discoveries to be made."

In practice, these three categories are not quite correct, since the first category as described has almost died out and there has been a degree of blurring between the surviving members of the first category and the second. In fact, we can probably divide works on punctuation more accurately into two broad categories: style guides and linguistic investigations.

3.1 Style Guides

Style guides are by far the most numerous type of works written about punctuation, and most of them tend to contain material that is identical, or similar to a large degree. The only differences tend to lie in the way that the material is presented, and in the degree of prescription that it is presented with.

There are several sub-types within this genre, which also need to be differentiated in terms of their time of production. The type of schoolteacherly lament described by Humphreys above is almost exclusively found in the older books (Husband and Husband, 1905; Skelton, 1949; Partridge, 1953; Carey, 1958). Funnily enough, the worst one of these is the latest (Carey, 1958) which seems to criticise and pontificate to such a degree that it is difficult, if not impossible, to extract the underlying messages about punctuation. This is perhaps due to the fact that it is actually a new edition of a much older work (Carey, 1939), which was perhaps more typical of its time. It is very rare for modern works on punctuation to do much in the way of editorialising — perhaps the only (mild) example is to be found in (Trask and Wale, 1993), who although they give a concise and informative guide to appropriate usage of punctuation, begin with a small amount of lamentation:

> "Why should you learn to punctuate properly? After all, many people have made successful careers without ever learning the difference between a colon and a semicolon."

Other modern works tend to be produced by publishing houses, associations with literary interests and for educational uses. Hence these tend more towards stark functionality without

any editorialising: they just describe the usage of punctuation marks for inexperienced users. In doing so, they manage to avoid the over-prescription of editorial Style Guides (Humphreys' second category), but are still rather prescriptive in nature, since they are aimed toward production rather than analysis. Examples of such more modern style guides are (McDermott, 1990; MHRA, 1991; Hilton and Hyder, 1992; Jarvie, 1992). Curiously enough, the two oldest style guides investigated (Anon, 1680; Allardyce, 1884) are comparable to the modern ones in terms of brevity and conciseness (presumably due to the desire for shorter, and therefore cheaper, books).

These style guides tend to contain roughly equivalent accounts of punctuation, although some are more prescriptive than others, and some seek to show all the possible uses of a particular mark. The punctuation functions they dicuss and permit are shown in table 3.1, where the two oldest style guides are also included for comparison and reference.

To explain some of the more opaque usage terms in table 3.1: the *critical* uses of the question and exclamation marks refer to those instances where authors seek to draw some form of rhetorical attention to the preceding word or syntactic entity, by introducing a degree of (sarcastic) emphasis (3.1) or uncertainty (3.2). With the comma, *delimiting* use refers to those instances where certain text fragments, such as relative clauses, are conceptually separated from the rest of the sentence, e.g. to make them non-restrictive (3.3); *disambiguating* usage refers to the cases where the comma is necessary to preserve the intended meaning of the sentence (3.4); *sentential inversion* relates to the use of a comma to prepose a text fragment to a sentence that would more usually follow it (3.5); and the usage of comma for *omission* refers to those cases where it marks the place that a word or words should have occurred in, usually in the case that it/they have already occurred in the sentence (3.6).

(3.1) He said he was enjoying (!) himself, but he didn't look happy to me.

(3.2) He said in rapid French that his name was Ellis (?) and that he wanted a meal — something like that.

(3.3) George IV, a great-grandson of Queen Victoria, died in 1952.

(3.4) She is sick, and tired of punctuation.

(3.5) Immediately she got home, she telephoned the police.

(3.6) Some savers prefer to invest their money in property; others, in the stock market.

The *connective* and *contrastive* uses of the semi-colon refer to the cases where that mark is used to join sentences together. The difference between the uses is determined by the nature of the adjacent sentences, as illustrated by the conjunctive sense (3.7) versus the contrastive one (3.8). The use of the colon in *book-titles* is illustrated in (3.9). The *discursive* use of the hyphen, the use of the apostrophe with certain *plurals*, and the principle of the *alternation* of quotation marks are illustrated by examples (3.10), (3.11) and (3.12) respectively.

(3.7) I will say no more; this matter is closed.

Punctuation	Uses	McD (1990)	MHRA (1991)	H & H (1992)	Jarvie (1992)	Anon (1680)	Allar. (1884)
dot	sentential	•	•	•	•	•	•
	abbreviations	•	•	•	•		•
	ellipsis (…)	•	•	•	•		•
capitals	sentential	•	•	•	•	•	•
	rhetorical (Oh, Wow)	•			•		
	proper nouns	•	•	•	•		•
	abbreviations	•	•	•	•		•
question mark	interrogatives	•		•	•	•	•
	within sentences						•
	critical (?)				•		
exclamation mark	emphasis	•		•	•	•	•
	within sentences						•
	critical (!)				•		
comma	lists	•	•	•	•	•	•
	adjectival lists	•		•	•		•
	delimiting	•		•	•		•
	disambiguating			•	•	•	•
	sentence inversion	•		•	•		•
	speech introduction	•		•	•		•
	omission				•		•
	numbers (17,000)	•		•	•		
	breath marker					•	
semicolon	lists	•		•	•		•
	connective	•		•	•	•	•
	contrastive			•			
colon	elaboration	•		•	•	•	•
	book titles				•		
	speech introduction	•		•	•		•
	sentential balance	•			•		•
	normal connective					•	•
dash	delimiting	•	•	•	•		•
	afterthought	•	•	•	•		
	elaboration	•		•	•		•
	elliptical					•	•
	breath pausing					•	•
hyphen	compounding	•	•	•	•	•	•
	special discursive	•			•		
apostrophe	possessive	•	•	•	•		•
	letter omission	•		•	•	•	•
	special plurals	•			•		•
quotes	quotation	•	•	•	•	•	•
	quote alternation	•	•		•		•
	'scare' quotes	•		•	•		•
brackets	parenthetical	•	•	•	•	•	•

Table 3.1: Comparison of punctuation coverage in various style guides

(3.8) You may be sorry; I am delighted.

(3.9) "Big Red Confidential: Inside Nebraska Football"

(3.10) Please speak s-l-o-w-l-y to me.

(3.11) Mind your p's and q's.

(3.12) "Fred shouted, 'Stop Thief!'," said John.

The other usage terms that may be unclear are some of those relating to the dash: usage to mark *afterthoughts* and *elaborations* involves a single dash separating a final text fragment from the rest of the sentence, as illustrated in (3.13) and (3.14), and under *elliptical* usage the dash is used similarly to the ellipsis mark to mark a discontinuity (3.15).

(3.13) Frank Harris is invited to all the great houses in England — once.
(attributed to Oscar Wilde)

(3.14) The Common Law of England has been laboriously built about a mythical figure — the figure of the 'Reasonable Man'. (A P Herbert)

(3.15) "Well, I'll be —," he muttered.

As can be seen from table 3.1, the two older style guides have rather different concepts of certain punctuation marks than the more modern ones do. However, much of the key functionality is preserved between the sets. Of the modern style guides examined, (MHRA, 1991) gives somewhat different coverage since it is a true Style Guide in the sense of Humphreys' (1993) second category. The only punctuation marks mentioned are those that could conceivably be used in an ambiguous or informal manner — all the more conventional uses are not mentioned.

A relation to the older, more 'schoolteachery' version of the style guides, and one which is not strictly-speaking designed to be educational except in a humourous and admonishing way, is the category of the occasional articles to be found (usually) in the popular press giving examples of recently observed punctuation misuses. Usually presented in an amusing style, these articles tend to focus most frequently on the 'sadly-abused' apostrophe — e.g. (Waterhouse, 1994) — and the misplaced comma — e.g. (Romm, 1984).

Then there are the true *Style Guides*, as presented in Humphreys' second category. Unlike the style guides discussed above, which correspond to at least some of the definition of Humphreys' first category (in that their purpose is to "teach punctuation to the peasantry") these *Style Guides* have more similarity to (MHRA, 1991). The distinction the two types is that the latter seek to force writers who are presumed to know already how to punctuate 'properly' just which marks they can use and in which circumstances, and that the former seek to teach people how to apply punctuation marks in general to achieve a comprehensible effect.

Since they presuppose a knowledge of the system of punctuation, the Style Guides are unlikely to be helpful to the current investigation. As illustrated by (MHRA, 1991), they rarely cover the whole spectrum of punctuation, and do little more than forbid particular uses of punctuation in a particular environment (books, theses, etc.) which are licensed in

more general usage. This type of document also occurs outside specifically editorially-based environments, as in (Leaver, 1992) which deals with document preparation in a legal context. Not only is it a *Style Guide* in the strict sense of the term, but it also expands on normal stylistic constrictions to try to elimate any possible ambiguities, which can of course be problematical in the drafting of legal documents.

The last category of stylistic guides to punctuation usage are the descriptions of the overall grammar and usage of the language, which often contain a chapter or section describing appropriate punctuation use (Fowler and Fowler, 1930; Partridge, 1954; Quirk et al., 1972; Quirk et al., 1985; Greenbaum and Quirk, 1990). The detail into which these go depends, of necessity, on the size of the overall grammar. In the case of (Quirk et al., 1972) the large size of the complete work means that some 30 pages can be dedicated solely to a discussion of punctuation, and so the treatment that punctuation receives in many of these works is as comprehensive as it would be in style-guide dedicated solely to punctuation.

All any style guide can describe, however, is the circumstances in which the use of a punctuation mark is appropriate. Whilst interesting for the formulation of a theory (to ensure that all such common situations are accounted for), these treatments are not sufficient to enable us to fully describe the function of the system of punctuation in a manner that would be relevant to a computational implementation.

In addition, by their very nature, style guides of all varieties are to some extent prescriptive, ranging from the harsh dictates of the publishers' *Style Guides* to the more gentle prescription inherent in the ordinary style guides. This prescription is necessary, to some extent, since these works are all intended to guide production of punctuation within written material. However, any computationally-applicable theory is likely to place a far greater emphasis on analysis rather than production, and so the account of punctuation that is being developed here should be a great deal more flexible and powerful than the prescriptive accounts of the style guides. It makes no sense to turn an analysis system into a determiner of punctuation correctness; an analysis that fails because of sloppy punctuation is of little use to any system. What a computationally-tractable account of punctuation should consist of is an account of the function of punctuation, so that any punctuation mark that is encountered in the text being analysed can be assigned as much information content as possible for the purposes of the ongoing analysis. Using any prescriptive notion of punctuation to guide such an analysis is likely to be counter-productive in the extreme, if not totally impossible.

However, a sufficient grounding for a computationally-tractable account of punctuation that avoids the over-prescription of the style-guide approach should be obtainable from those works that are aimed more at describing the function and operation of the whole system of punctuation, rather than at the usage of individual marks. These works are those that Humphreys (1993) refers to in his third category, namely those that address the linguistics of punctuation.

3.2 Linguistic Treatments of Punctuation

As Humphreys mentions (1993), there are relatively few texts that have addressed in any detail the function of punctuation as a linguistic system. The reason for this is that from the point of view of theoretical linguistics, the developments of the punctuational system in recent centuries appear to have been overlooked. It has only been relatively recently that punctuation has been recognised as an orthographic system in its own right that does not simply transcribe pauses in speech and other intonational features (or that does not transcribe these for *silent* reading) (Nunberg, 1990).

In the description of punctuation in their English Grammar, Quirk et al. (1985) describe a hierarchy that relates structural punctuation marks to the particular grammatical units that they can separate or enclose. Table 3.2 illustrates the normal positions on the hierarchy that the punctuation marks occupy, but it is acknowledged that many of the marks can occupy other positions on the hierarchy, but usually only do so rarely and for particular stylistic effects.

Building on the hierarchy in table 3.2, Meyer (1987) was one of the first to devote serious study solely to the issue of punctuation. In an analysis of the Brown corpus of American English he examines the uses that the various marks of punctuation are put to, and the circumstances in which they are used. Drawing on this information, and synthesising it with the more prescriptive accounts found in style guides, Meyer postulates a set of rules and principles for American punctuation function.

Meyer regards punctuation rules as specifying the permissible places within a text where marks of punctuation can be placed, and principles as specifying which of the punctuation marks are most appropriate in those contexts. In sentence (3.16), for example, to account for the placement of a comma before *and* requires a rule stating that compound sentences can be conjoined by a comma, semicolon, dash or full-stop, and a principle stating that a comma is typically used to punctuate the clauses of a compound sentence if those clauses are short, non-complex and conjoined by *and*.

(3.16) Paul Paray, rounding out his current stint with the orchestra, is a solid musician, and the Philharmonic plays for him. (Meyer, 1987)

	Mark	Grammatical Unit Set Off by Mark
Level 1	Full-stop Question Mark Exclamation Mark	Sentence — " — — " —
Level 2	Colon Parentheses Dash	Sentence, Clause or Phrase — " — — " —
Level 3	Semi-colon	Clause Clause or Phrase in a series
Level 4	Comma	Clause or Phrase

Table 3.2: Quirk et al.'s Hierarchy of Punctuation Marks

Meyer produces a 'rule-table' of 13 syntactic situations where punctuation is permitted and indicates the marks that are permitted there, separating them into common and uncommon usages. In addition he produces 8 principles — syntactic, semantic and prosodic — which determine when the positions described by the punctuation rules are filled, and with which marks they are filled.

The drawbacks of Meyer's research is that it is specifically based on American English, whose norms and practices are possibly divergent from British English and certainly so for other languages. In addition, the resulting principles and rules are once again rather production oriented. It would be non-straightforward to reverse their sense of application and use them for analysis. Also, once again it is rather the case that the positioning and usage of punctuation marks has been described (albeit in a far more useful, non-prescriptive way), rather than an account of how the whole system of punctuation works.

The research that addresses this issue, and perhaps the most influential piece of work in the punctuation field so far, is found in (Nunberg, 1990)

Nunberg's Account of Punctuation

Nunberg, considering only those marks of punctuation that occur below paragraph level (which therefore correspond to the previous definition of *inter-lexical punctuation*) grounds his theory of punctuation by distinguishing between two different levels of grammar of the written language. The first he describes as the *lexical* grammar, which is reponsible for the description of the relationships and dependencies that hold between the lexical items in the text (e.g. words). In addition to the lexical grammar, however, Nunberg also postulates the existence of a *text* grammar for the rules that describe the distribution of explicitly-marked non-lexical categories such as the paragraph, text-sentence and dash-interpolation, which classify the role of the content of the lexical constituents relative to the overall argument and the context of interpretation.

Thus under this theory, punctuation can be divorced slightly from the lexical items it comes between, and indeed Nunberg argues that to a certain extent the text-grammar can be determined and applied independently of the content of the lexical categories.

Within his text-grammar, Nunberg defines further categories such as text-clause and text-phrase (and, as we have seen above, text-sentence). Thus a paragraph consists of one or more text sentences, and a text sentence consists of one or more text clauses, which are usually separated by semicolons (3.17).[1]

(3.17) John took French.
John took French; Susan took Spanish.
John took French; Susan took Spanish; Annette took Russian.

The constituency of the text clause is a more problematical matter. Nunberg's first restriction on text clauses is that they cannot simply recurse as expansions of themselves. Thus while (3.18) is valid, (3.19) is not.

[1] Examples in this section are taken from (Nunberg, 1990) unless otherwise indicated.

(3.18) The students were allowed to choose which language they wanted to study; Jan took Spanish.

(3.19) *The students were allowed to choose which language they wanted to study; Jan took Spanish; Betty took French.

Text clauses are further augmented by the introduction of the category of clausal adjuncts. This category includes colon expansions (3.20), dash interpolations (3.21) and literal parentheticals (3.22). The colon expansion is restricted to being the right-most element of the text clause that contains it, while the other two types of clausal adjuncts can occur anywhere within the text clause except at the beginning, with certain other provisos, for example that they cannot occur adjacent to one another at the same level (3.23).

(3.20) I will be frank: there is no way you're going to get the job.

(3.21) And — what's more surprising — she left.

(3.22) And (not surprisingly), she left.

(3.23) *She walked out — who could blame her — (it was during the chainsaw scene, as I recall) and went directly home.

The adjuncts themselves are unconstrained as to their content, except that it must be below text sentence level. The single exception to this is with the parenthetical construction. Since it is not restricted to its role of clausal adjunct and can actually occur at any level in the text, the parenthetical construct could actually contain several paragraphs or sentences, depending on the level it occurred at. However, if the parenthetical has been started as a text-clausal adjunct, below sentence level, it too is constrained to include only text categories below the sentential level. The content of clausal adjuncts can therefore be defined as one or more text clauses, with the proviso that they should not directly nest, i.e. a colon expansion should not contain another colon expansion (3.24); a dash interpolation should not contain another, at least at the top level (3.25); and a parenthetical should not contain another parenthetical expression (3.26), unless it uses different parentheses ([{}]) or the internal parenthetical is a special case (a bibliographic reference, for example).

(3.24) *They serve a lot of cajun dishes: blackened redfish, gumbo, and one thing you don't see a lot of: catfish sushi.

(3.25) *Holding the bat firmly — her father — himself a former major leaguer — had given it to her — she awaited the pitch.

(3.26) *Jones (an employee (actually, a director) of the firm) was also present.

An interesting result of these principles is the possible ambiguity of the colon expansion in those cases where it contains multiple text clauses. Since the semi-colon in examples (3.27) could have either wide or narrow scope, as illustrated in (3.28), the sentence is ambiguous. Since colon expansions are prevented from containing further colon expansions, however, the sentence in (3.29) is unambiguous, with the semi-colon having wide scope (3.30).

(3.27) The press secretary gave them the rules: they were not allowed to speak to the committee directly; all other members were forbidden to discuss what the committee had decided.

(3.28) **[** [words **:** words **]** **;** words **.** **]** *wide scoping*
[words **:** **[** words **;** words **]** **.** **]** *narrow scoping*

(3.29) The press secretary gave them the rules: they were not allowed to speak to the committee directly; all other members were forbidden to discuss what the committee had decided: a hiring freeze would take place.

(3.30) **[** [words **:** words **]** **;** **[** words **:** words **]** **.** **]** *wide scoping*

Nunberg additionally provides a set of rewrite rules for this part of his text grammar to represent the structure of the text-categories he refers to, which in effect restate formally the principles discussed above:

$\mathcal{S}_t \Rightarrow \mathcal{C}_t^+$ (this one not actually given in (Nunberg, 1990))
$\mathcal{C}_t \Rightarrow \mathcal{P}_t^+(\mathcal{X}_c)$
$\mathcal{X}_c \Rightarrow \mathcal{C}_t'^+$
$\mathcal{P}_t \Rightarrow \mathcal{E}(\mathcal{A}_\mathcal{C})$
$\mathcal{A}_c \Rightarrow \mathcal{C}_t'^+$

where

$\mathcal{S}_t$ is a text sentence, $\mathcal{C}_t$ is a text clause, $\mathcal{P}_t$ is a text phrase, $\mathcal{E}$ is a lexical expression, $\mathcal{X}_c$ is a colon expansion, $\mathcal{A}_c$ is the other clausal adjuncts, and $\mathcal{C}_t'$ is a restricted text clause w.r.t. the clausal adjuncts it can contain.

The comma is introduced by Nunberg in relation to the interaction of the text and lexical grammars. He divides instances of the comma into three categories: delimiting commas that go *around* a lexical element (3.31), separating commas that go between them (3.32) and disambiguating commas, that Nunberg mentions only in a footnote, which separate items of different syntactic types to prevent ambiguity or parsing difficulties (3.33).

(3.31) The key, which I did not have duplicated, has been lost.

(3.32) The woods are lovely, dark and deep.

(3.33) Those who can, contribute to the fund.

It is suggested that while separating commas occur only at the level of the lexical grammar, delimiting commas have relevance both to the text and lexical grammars, since they seem to have some similar text-grammatical functions to text-clausal adjuncts: a comma-delimited phrase cannot begin with the dash and parenthetical clausal adjuncts (3.34), whilst a comma-separated phrase can (3.35,3.36).

(3.34) *The bombings, (then Secretary of State) Henry kissinger announced, were an important step towards peace.

(3.35) Among those present were Le Duc Tho, (then Secretary of State) Henry Kissinger, and the French foreign minister.

(3.36) But proponents of this new, (logical) positivist view could not countenance the introduction of such entities.

It is important now to distinguish Nunberg's *text grammar* from the notion of a *punctuation grammar*. The text grammar has proved not to be a grammar that deals with punctuation marks. Some of the higher level phenomena handled by the text grammar coincide with punctuation marks, but, as exemplified by the case above of the separating comma, not all punctuation marks are relevant to the text grammar. It is possible to take this differentiation further by examining the special principles that Nunberg postulates for the interaction of punctuation marks.

The semi-colon is a mark that so far in Nunberg's theory has appeared exclusively to separate text clauses, and hence is intimately connected with the text grammar. However, the semi-colon also has another use that falls exclusively into the realm of the lexical grammar, namely when it replaces commas by the principle of **semi-colon promotion**. Nunberg formalises a rule that requires that when items containing commas are conjoined, the separating commas that would usually perform the conjoining task in these instances are replaced (or 'promoted to') semi-colons (3.37–3.40) (where I have used ∼ to represent an interpretational ambiguity). This rule of semi-colon promotion requires that the lexical conjunction before the final item is preceded by a semi-colon, even if there would not have been a comma there. In addition, the rule does not apply when only the last item of the conjoined list contains the comma (3.41), and only ever applies on the highest level separators (3.42) despite any ambiguity. The example (3.42) refers to three books, two about a single baseball player, and one about three.

(3.37) Among the speakers were John; Ed; Rachel, a linguist; and Shirley.

(3.38) ∼ Among the speakers were John, Ed, Rachel, a linguist, and Shirley.

(3.39) Rachel will chair the first session, and the second session will be postponed; or I will chair both sessions.

(3.40) ∼ Rachel will chair the first session, and the second session will be postponed, or I will chair both sessions.

(3.41) Among the speakers will be John, Ed, and Rachel, a linguist.

(3.42) He has written books on Babe Ruth; on Tinker, the shortstop, Evans, the second baseman, and Chance; and on Hank Aaron.

Nunberg's other principles govern the interaction of adjacent marks of punctuation, and can be divided into two categories: transposition rules that exchange the positions of two

marks of punctuation, and absorption rules that remove one of the punctuation marks, leaving only a single mark behind.

Transposition

There is only really one member of the transposition category, namely the principle of **quote transposition**. This requires that certain marks of point punctuation (those shown in (3.43)) that occur immediately to the right of a closing quotation mark (i.e. outside it) are moved to the immediate left of that mark (3.44–3.46). This rule can also apply recursively in the cases where several closing quotation marks occur together (3.47).

Of course, the quote transposition rule is based on standard practice in American English punctuation. In British English, the possibility exists of not transposing the punctuation marks, depending on whether the quoted matter is making the 'primary assertion' of the sentence. Thus (3.48) is untransposed, whereas (3.49) is transposed. It is this flexibility which would explain the non-transposition of the other point punctuation marks, colon and semi-colon, and would prevent transposition from happening with 'scare quotes'.

(3.43) `. ? ! ,`

(3.44) Nelson actually said, "England *confides* that every man will do his duty."

(3.45) Nelson actually said, "England *confides* that every man will do his duty," but the message was changed by the signal officer.

(3.46) Nelson actually said, "England *confides* that every man will do his duty"; but the message was changed by the signal officer.

(3.47) Then the Lord said unto moses: "Go in unto Pharoah, and tell him: 'Thus saith the Lord, the God of the Hebrews: "Let my people go, that they may serve me."'"

(3.48) Some people insist that Nelson's last words were really "Kismet, Hardy".

(3.49) "This policy," he said, "will bring the government to ruin."

The driving force behind the practice that Nunberg formalises in this rule is derived from typesetting. The white-space underneath the final quotation marks is seen as disrupting the natural reading movement of the eye, when the final punctuation mark still needs to be registered (3.50), especially relevant in those cases where double quotation marks are composed from two separate inverted commas (apostrophes), when the whole quotation mark is wider than it appears here. With quote transposition, on the other hand, the end of the sentence is nicely tapered, since the width of the dot or comma is not sufficient for the white-space above it to disrupt eye-movement to the quotation mark (3.51).

(3.50) "Quoted text".

(3.51) "Quoted text."

Absorption

Nunberg's other principles cover the absorption of one punctuation mark by another adjacent mark. The main rule is the one covering **point absorption**. When two or more point symbols (3.52) occur adjacently, they are absorbed according to a power hierarchy (3.53). In addition, multiple instances of the same punctuation mark will be self-absorbed to leave just a single instance of that mark.

(3.52) . ; : — ,

(3.53) . ≫ ; : ≫ — ≫ ,

Thus in (3.54), absorption has occurred between two final delimiting commas to leave a single one; in (3.55) a final delimiting comma is absorbed by a semi-colon; and in (3.56) a final delimiting dash is absorbed by a full-stop. It is not the case, however, that these absorptions are due to the superiority of marks which occur at a higher textual level: in (3.57) it is the final delimiting comma that is at a higher textual level to the final dash, but the comma is still absorbed according to the hierarchy in (3.53).

(3.54) Hagy, who resigned in 1985, in fact, protested the policy.

(3.55) John left, apparently; Mary stayed.

(3.56) We heard the sound of their artillery — devastating in its fury.

(3.57) I am glad you asked me that, my friends — if I may call you that — because I have a good answer.

As an extension of point absorption, the principle of **bracket absorption** governs the interaction of point punctuation with parenthetical marks and quotation marks. The principle requires that when a mark of point punctuation occurs immediately inside (to the left of) a final bracketing character, it is absorbed by that bracketing character. Thus in (3.58) a final delimiting comma is absorbed by the right parenthesis, in (3.59) a dash is similarly absorbed; and in (3.60) a final dash is absorbed inside a final quotation mark. Marks that naturally occur to the outside of such a bracketing character are left unaltered (3.61). Of course care must be taken regarding the possible interaction of this principle of bracket absorption and that of quote transposition, and a hierarchy of application must apply — bracket absorption occurring before quote transposition — so that transposed material is not immediately absorbed. In (3.62), the final dash is first absorbed before the full-stop is transposed.

(3.58) May failed the test (she had not studied the material, which was handed out when she was absent) and will have to repeat the course.

(3.59) May failed the test (which was not surprising — she didn't study) and will have to repeat the course.

(3.60) David announced that "the test will be a breeze — you don't have to study" and went out to play ball.

(3.61) May failed the test (given on Monday), which covered all of the readings, and will have to repeat the course.

(3.62) David announced that "the test will be a breeze — you don't have to study." *(not an example from (Nunberg, 1990))*

The final principle in this category is that of **graphic absorption**. This principle requires the absorption of the full-stop by orthographically (although not linguistically) similar symbols, such as the abbreviation-marking point (3.63) or the stress-markers (3.64). Note that these are not treated as conventional point punctuation since they fail to absorb other non-point punctuation marks (3.64) and graphically dis-similar marks of point punctuation that they occur adjacent to (3.65). As a further result, the stress markers and the abbreviative point are not absorbed by bracket absorption either.

(3.63) He lives in Washington, D.C.

(3.64) Are you going to Washington, D.C.?

(3.65) It requires a sentence period; i.e., a full stop.
What we want to know is when?, but he has only told us why.

As discussed already with respect to bracket absorption and quote transposition, these principles have the capacity to interfere with one another. Therefore Nunberg additionally postulates an ordering on these rules as follows, where *realisation* rules are the rules that convert abstract syntactic categories into instantiations of punctuation marks (e.g. from a non-restrictive clause into a set of lexical items delimited with commas) — *semi-colon promotion* therefore is a principle that operates at the level of the realisation rules:

realisation rules
⇓
point absorption
⇓
bracket absorption
⇓
quote transposition
⇓
graphic absorption

Nunberg then goes on to consider **pouring rules**, which relate to the way that the text is laid out on the page (paragraph breaks, line breaks, etc.), as well as the functions and usages of the individual punctuation marks (much the same as presented in style guides). Therefore the main contribution of the work in (Nunberg, 1990) for the current investigation is the notion of text-grammar and the formalisation of the various principles that apply to punctuation marks when they interact.

3.3 Computational Approaches

As mentioned in chapter 1, computational treatments of punctuation phenomena in natural language processing systems are still relatively rare, except for the full-stop which is often used to recognise sentence breaks, although even this utilisation has associated difficulties and ambiguities (Palmer and Hearst, 1994). Many systems simply strip punctuation marks other than the full-stop from text to be analysed or do not include such marks in generated text. The systems which do try to make use of punctuation marks tend to do so in rather ad-hoc fashions which do not account for all instances and meanings of punctuation marks and have little in the way of theoretical motivation.

In the TACITUS system for message understanding, for example, punctuation marks are used chiefly for segmentation (Hobbs, 1991). After regularisation by a pre-processor (which is already likely to destroy some of the information-content of the punctuation marks as used originally) the punctuation marks in a message are used to segment messages into sentences. Commas are then utilised to further segment long sentences before they are parsed using the specially developed *terminal substring* parsing method, whereby commas, conjunctions and certain other lexical phenomena are used to subdivide sentences into chunks which are more easily and rapidly analysed than the whole sentence. Certain instances of comma placement are incompatible with this type of segmentation, and so cause problems for the analysis process. In addition, several other types of punctuation phenomena are not catered for and hence are problematical to the analysis stage: these include parentheses, dash-interpolations and colon-expansions. Since the treatment of punctuation occurs at the syntactic stage, the above problems are acknowledged by Hobbs as shortcomings in the grammar. This serves to illustrate how the marks that have been catered for have been implemented on an ad-hoc basis to meet the most immediate requirements of the system, rather than in any more principled manner that reflects a theory of punctuation.

A similar approach is taken in the CLE system (Alshawi, 1992). When the system is used for analysing text, punctuation marks are used in the pre-processing stages to segment long sentences into manageable chunks. The marks used for this purpose are commas, semi-colons and, interestingly, list indicators which are present in the SGML mark-up of the text. These segments are then processed individually before the results are reassembled at the end of the analysis. In addition to this segmentation role, certain punctuation marks are also included in the analytical grammar. Commas are included to facilitate the analysis of complex conjunctive phrases (where their semantics is picked up by unification with the final, lexical, conjunct), and matched parentheses and dashes are also catered for. Other instances of punctuation are treated via the *special form* mechanism whereby new lexical entires are created on the fly for unknown items. All punctuation is treated via regular expressions in the CLE, although the desirability of including a statistical component is acknowledged. Thus whilst several punctuational phenomena are addressed by this system, the choice and implementation of these is still rather unprincipled and driven by considerations of system performance and complexity.

3.4 Comment

Most of the theoretical works investigated, whether style guides or more linguistic works, give good definitions and examples of the usage of individual punctuation marks, and there is a great deal of correlation between the definitions provided by the various works. Punctuation is currently treated in only a few computational systems, and those treatments are at best partial and motivated more by simple performance enhancements than by any principled account. To make these accounts more complete and well-motivated, a theory of punctuation function needs to be synthesised that can be easily implemented in computational systems.

However, for the purposes of the formulation of a computationally-applicable theory of punctuation, more is needed than prescriptive usage definitions of the various punctuation marks involved: it is necessary to find out how the system of punctuation operates as a whole, within normal, lexical text. The most relevant work, therefore, for the formulation of a theory, is that in (Nunberg, 1990), which provides a solid background for further investigation.

The concept of the text-grammar is an interesting and valuable one, but as we have seen it is not necessarily directly relevant to the usage and analysis of punctuation. Indeed, there are problems associated with the use of a text grammar, and its integration in parsing with the lexical grammar, which will be discussed in later chapters.

The principles guiding the interaction of punctuation marks, however, are particularly valuable, and crucial to any treatment of punctuation, since they provide a mechanism for accessing punctuation information not present at the surface orthographic level (final delimiters, for example, if absorbed).

four Justification for Investigating Punctuation

"I live across the street — My name is Marcie — Marcie Dahlgren-Frost: Dahlgren is my maiden name, Frost is my married name. I'm single again but I never bothered to lose the Frost. And I get compliments on the hyphen!" (Laurie Metcalf, "Uncle Buck")

Realisation of the importance of punctuation in the production and understanding of written text, and the study of its development and changing roles over time, are very interesting and useful but unfortunately these factors are not enough to stimulate a computational treatment of punctuation. Although it is acknowledged that punctuation has an important role to play in written language, it is not necessarily obvious that a computational system would be able to make profitable use of a treatment of punctuation. Of course, such consideration should bring improvements (since we have seen what a crucial role punctuation has to play in the written language), but these may be outweighed by disadvantages: it could be the case that introduction of punctuation into a language processing system introduces huge computational overheads, for example, and makes efficient processing intractable. The question to be answered here, then, is whether a theory of punctuation is likely to work advantageously in a computational context.

Since this global 'computational context' consists of many facets — generation, understanding, semantics, parsing etc. — it makes sense to examine the efficacy of punctuation in some of these facets separately, rather than in some universal manner. One justificational study, that will be described here, has been carried out in the fields of natural language generation and discourse processing. The other major area that it seems could benefit from a justificational study is the area of syntax and parsing, and the issue of whether consideration of punctuation can assist parsing results and efficiency will also be addressed later in this chapter.

4.1 Generation and Discourse Structure

Robert Dale's paper (1991) on the role of punctuation in discourse structure is motivated by the view that "natural language processing systems, whether generating or understanding language, should take account of the physical manifestation of language: in the context of written language, this means taking account of the constraints imposed and opportunities offered by the real estate of the pages (or whatever other surface) on which the words appear." An obvious and natural corollary of such a view is to take account of punctuation in the written language.

Dale takes the view that, in many uses of punctuation, the marks act as signals of discourse structure, and whilst he acknowledges that some of these uses are amenable to straightforward syntactic characterisation (separating commas between items in a list, for example) he concentrates on the semantic aspects of punctuation use. Many argue against the clarification of the role of punctuation in indicating discourse structure, citing the apparent non-determinism of punctuation use and idiosyncrasies in use between different writers. Dale however argues that the disagreements in punctuation use are more about the strength of punctuation appropriate in a given circumstance (the use of a comma rather than a semi-colon, for example) rather than disagreements based on conflicting use.

He argues that rather than there being no rules, these do exist for the application of punctuation, but they are learnt and applied differently to the conventional rules of grammar. Since punctuation only occurs in written language, people are less practiced in its use than for the rules of grammar proper, and hence regard any rules for punctuation as artificially stipulated conventions. Publishers' style guides, one of the classes of the small number of works about punctuation, to a certain degree give these 'rules' of punctuation when they describe the 'correct' use of a punctuation mark and the contexts it can be used in.

The argument for some amount of determinism in the use of punctuation is reinforced through the style guides — they all take the view that there are clear correct and incorrect ways to use punctuation, and that the punctuation marks should be used to make clear certain relationships in the text. It is important to appreciate, however, that these two sets do not necessarily together define the universe of punctuation use — it seems more than likely that there is a fuzzy area in the middle where idiosyncrasy and disagreement reign supreme. Furthermore, since there is a remarkable degree of correlation between the various guides, it seems that the guidelines given in each are not merely arbitrary or idiosyncratic to the particular author or editor. The problems with the definitions given in the style guides is that they tend to be insufficiently-well specified to be of use in any computational context.

Dale then goes on to examine some examples of the use of punctuation, considering the natural interpretations of such uses with respect to key notions in discourse structure theories such as Rhetorical Structure Theory (Mann and Thompson, 1986) and Grosz & Sidner's Discourse Structure Theory (1987).

For example, the sentences given in (4.1) — (4.3), taken from (Nunberg, 1990), each seem to embody the same two basic propositions, shown in (4.4).

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

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

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

(4.4) P_1: Some person reported the decision.
P_2: We were forbidden to speak with the chairman directly.

Since the readings of the three sentences are different, it seems plausible that any differences in perceived meaning are due to the different punctuation marks in the separating of the two propositions. In the terms of Rhetorical Structure Theory, for example, examples (4.1) and (4.2) seem to exhibit instances of the ELABORATION and CAUSE relations respectively, whilst (4.3) could express either of these rhetorical relations: that is that in (4.1) the fact of P_2 seems to be the content of the decision, whereas in (4.2) P_2 is the reason for P_1, or at least the reason for someone other than *the chairman* reporting *the decision*. Note that this is Dale's analysis — other readers may get slightly different flavours of the meanings of these sentences.

Underdetermination

Dale argues that it should "not be particularly surprising that the particular rhetorical relations that reside in a text are underdetermined by the punctuation marks that are used…[T]he situation…is not all that different in kind to that which happens in the case of lexical markers, since the same cue words can be used for a number of different rhetorical relations."

The question arises as to the precise nature of the underdetermination. If a taxonomy of rhetorical relations is adopted, as in (Hovy, 1990), the use of a particular punctuational marker might be rooted at a particular node in the taxonomic tree, with all the rhetorical relations that are descendants of that node being able to employ that particular punctuational marker as a clue to the relation being expressed. Similarly, and vice-versa, the distribution of punctuational markers might tell us something about how the space of rhetorical relations should be structured.

Some of the punctuation marks are so lacking in specificity, however, that they do not provide much information as to the relevant rhetorical relation. Borrowing again from (Nunberg, 1990), example (4.5) is such a non-specific case.

(4.5) Some people found the book fatuous; John considered it a paramount example of postmodern criticism.

Depending on context, the relevant rhetorical relation here could be ELABORATION — John's view of the book is an example of the general sentiment in the first proposition — or one of ANTITHESIS — John's view is being contrasted with the general one. A given reader's interpretation of the punctuation mark, if indeed it is playing any role at all, will be strongly influenced by that reader's view of postmodern criticism.

Thus the possibility exists that if the use of punctuation is to be rooted in a taxonomy of relations, the range of relations characterised by each mark may be so broad that the marks have to reside in the least specific nodes of the taxonomy, and hence the rooting loses a great deal of its value.

Discourse Structure and Grain Size

A further question when considering punctuation from the point of view of discourse arises with the determination of just what units are being related. Text spans and discourse segments are almost invariably of clausal size or greater. Some punctuation, however, suggests that it might be necessary to look within individual clauses when attributing discourse structural relations to a text. Sentences (4.6) and (4.7) convey the same information, but since discourse theories have little to say about sentence-internal phenomena this suggests that if punctuation is to play a role then new theories must take account of such clause-internal material.

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

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

Therefore there is an argument for viewing the use of punctuation marks as simply one possible realisation of hierarchical structural relations that could be realised in other ways (for example, a more graphical process of line breaking and indentation), and that all these relations influence the discourse structure. Hence discourse structure below sentence level is a reality, which the prior emphasis on spoken material has deceptively missed.

Towards a Taxonomy of Punctuation Function

Dale suggests that punctuation marks can indicate:

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

The first deals with distinction between NUCLEUS material and SATELLITE material in a written text. The satellite material is that less important material in a sentence that can be deleted without the sense of the text being affected, whereas if the nucleus material is deleted the text becomes incoherent. It seems possible, even probable, that this distinction can be marked with punctuation, e.g. parentheses or the dash interpolation.

The second point, that of aggregation, describes that function of punctuation whereby the marks can indicate how closely related the material is. A comma suggests a closer relationship than a semi-colon for example, and differing degrees of embedding in a sentence, as manifested by the delimiting punctuation marks suggest more distant, tenuous relationships.

The third point, regarding the specification by punctuation of particular rhetorical relations, seems to be very clear. Style guides are generally fairly unanimous on the fact that colons are used to emphasise a sequence in thought or to illustrate, amplify or explain (4.8). Similarly commas, according to some style guides, may be used to emphasise a point or subtle distinction or to indicate contrast (4.9).

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

(4.9) As there were wicket-keepers before Blackham, so there were labour unions before the gold discoverers.
The fool wonders, the wise man asks.

Such categorisation is acknowledged to be "very impressionistic" and it remains to be seen whether the uses of punctuation can be tied to the theoretical notions underlying current theories of discourse structure in such a straightforward manner.

Summary

From these investigations, Dale concludes that punctuation is of very definite use in the field of discourse structure, but suggests further research into the relationship between punctuation and lexical markers, the mapping from punctuation onto the taxonomy of relations, the choices between graphical, punctuation and lexical markers and the nature of an atomic unit of discourse structure and the boundary between discourse and syntax.

4.2 Understanding and Syntax

Building on the recognition that punctuation has an important role to play in discourse structure, semantics and natural language generation it is useful to examine the complementary field, namely that of natural language understanding (or at least analysis) and in particular, syntax. The following experiment explores the role of punctuation marks in assisting the parsing of text, by constructing a grammar which is used to parse a small set of varied test sentences. The parse performance of this grammar is compared with that of an identical grammar constructed by taking the punctuated grammar and stripping out all references to punctuation marks.

The Grammar

The aim of this experiment is to discover whether the inclusion of punctuation into a natural language grammar can improve parsing performance, but first it is important to realise that recognition of punctuational phenomena does not imply that they can be successfully encoded into a natural language grammar.

Nunberg (1990) advocates the use of two separate grammars, operating at different levels. A lexical grammar is proposed for the lexical expressions occurring between punctuation marks, and a text grammar is proposed for the rough structure of the text, including most of the punctuation phenomena and an indication of the relationship of those phenomena to the lexical expressions adjacent to them. The text grammar has within it distinct levels, such as phrasal and clausal, at which distinct punctuational phenomena can occur.

This should, in theory, make for a very neat system: the lexical syntactic processes being kept separate from those that handle punctuation. However, in practice, this system

seems unlikely to succeed since in order to work, the lexical expressions that occur between punctuation marks would have to carry additional information about the syntactic categories occurring at their edges so that the text grammar can constrain the function of the punctuation marks.

For example, if a sentence includes an itemised noun phrase (4.10), the lexical expression before the comma must be marked as ending with a noun phrase, and the lexical expression after the comma must be marked as starting with a noun phrase. A rule in the text grammar could then process the separating comma as it clearly comes between two similar syntactic elements.

(4.10) I am seeing William, Andrew and Tom.
[end: np] [start: np]

However, as (4.11) shows, the separating comma concept could require information about the categories at arbitrarily deep levels occurring at the ends of lexical expressions surrounding punctuation marks.

(4.11) I like to walk, skip, and run.
I like to walk, to skip, and to run.
I like to walk, like to skip, but hate to run.

In addition to the problems of correctly processing a punctuation mark, thought must be given to the probable resulting linguistic analysis of the whole sentence. If lexical parsing were to occur before the punctuation stage, then some form of grammar would have to be used that permits incomplete analyses (e.g. transitive-verb-phrase-requiring-object, or some notion of optional additional constituents, so that any syntactic category occurring at the right-hand end of the complete syntactic unit can reappear after separating punctuation marks). In the sentences in (4.11), for example, the perfectly grammatical sentence *I like to walk* must be marked for the possibility of extra verbal objects. The only alternative to this is not to carry out a complete parse of the lexical expressions around the punctuation, and to unify lexical and punctuational components before carrying out the final parse.

Therefore even with the edge-category information discussed above, the parsing process is not necessarily made any easier (since often the full partial parses of all the separate expressions will have to be held from the lexical stage and joined at the punctuation stage). Therefore we seem to be at no advantage if we use this approach. In addition, it is difficult to imagine what linguistic or psychological motivation such a separation of punctuation from lexical text could hold, since it seems rather unlikely that people process punctuation at a separate level to the text it surrounds.

Additionally, since Nunberg advocates that the text and lexical grammars function separately to one another, it would be impossible for the text grammar to be able to differentiate the various uses of the comma since these are usually only apparent from syntactic context.

Hence it seems more sensible to use an integrated grammar, which handles both words and punctuation. This lets us describe the interaction of punctuation and lexical expressions far more logically and concisely than if the two were separated. Good examples of this are

Nunberg's so-called *disambiguating* commas (4.12); in a unified grammar we can simply write rules with an optional comma among the daughters to mark that a comma could occur there if necessary (4.13).

(4.12) Those who can, contribute to the fund.
She was sick, and tired of punctuation.
Such women as you, are seldom troubled with remorse.

(4.13) s → np (comma) vp.
s → pp (comma) s.
vp → vp (comma) vp.

A feature-based grammar of extended part-of-speech tags was written for this investigation (based loosely on one used in (Briscoe and Waegner, 1992)), and used in conjunction with the parser included in the Alvey Tools' Grammar Development Environment (GDE) (Briscoe et al., 1987; Carroll et al., 1991), which allows for rapid prototyping and easy analysis of parses. It should be stressed that this grammar is solely one of tags, and so is not very detailed syntactically.

The principal modification to the grammar to specifically handle punctuation was the introduction of the notion of the *stoppedness* of a category. Through this, every category in the grammar has a **stop** feature which describes the punctuational character following it (4.14), and defaults to **[st –]** (unstopped) if there is no such character. Thus punctuation marks are in effect treated as clitic on words, necessitating the addition of extra featural information to the rules of the grammar.

(4.14) the man, ⇒ np[st c]
with the flowers. ⇒ pp[st f]

As a corollary of the stopping principle therefore, the rules of the grammar further need to dictate that the mother category inherits the stop value of its rightmost daughter, and hence only rules to specifically add punctuation for categories which could be lexicalised are necessary. Thus a rule for the addition of a punctuation mark after a lexicalised noun would be as in (4.15). (The GDE terminology for a unification variable is the @-sign followed by an identifying letter.)

(4.15) n0[st @s] ⇒ n0[st –] [punc @s]

It is then straightforward to specify that top level categories must end with a full stop **[st f]**, that items in a list should be **[st c]** (comma), etc. In rules where it is necessary to force a particular punctuation mark to the right of a category, that mark can be included in the rule, with the preceding category unstopped: (4.16) illustrates the addition of a comma-delimited noun phrase to a normal noun phrase. Specifically mentioning the punctuation mark prevents the possibility of the mother category (i.e. the overall noun phrase) being unstopped, since a requirement of the delimitation is that the delimiting phrase ends with a comma, or, through point absorption, with another punctuation mark. Note that the phenomenon of point

absorption has therefore been captured by unifying the value of the *stop* feature of the mother and the identity of the final punctuation mark. Thus in processing the examples in (4.17) with the rule in (4.16) the delimited phrase *Bill Clinton* is always unstopped, and the stop value of the whole noun phrase (shown alongside the sample sentences) will therefore be whatever the punctuation mark following the delimited phrase is. Thus the possible values of **st** are all the possible values of **punc** in addition to **[st –]**.

(4.16) np[st @s] ⇒ np[st c] np[st –] [punc @s].

(4.17)

The President, *Bill Clinton* **,** entered.	np[st c]
The President, *Bill Clinton* **;** the Queen and…	np[st sc]
More news about the President, *Bill Clinton* **:** he's…	np[st co]
…it's the President, *Bill Clinton* **.**	np[st f]

In a similar fashion the disambiguating (optional) comma as discussed above (4.13) can be introduced by specifically mentioning it in a rule, but stipulating that the preceding daughter-category of the rule must be unstopped, i.e. **[st –]**, as in (4.18). Although this usage of a comma between the subject noun phrase and predicate verb phrase seems wrong, it does appear valid for disambiguation in certain circumstances as in the example in (4.19).

(4.18) s[st @s] ⇒ np[st –] ([punc c]) vp[st @s].

(4.19) Those who can, contribute to the fund.

An interesting problem associated with this method of inserting punctuation is that of the bracket characters. As discussed in section 3.2, the closing bracket characters (quotation marks and parentheticals) absorb internal point punctuation, but are able to co-occur with external point punctuation. In the case of quotation marks, the point punctuation can further be transposed inside the closing mark. Therefore sequential constructions such as *quote comma*, or *close-parenthesis full-stop* are perfectly legitimate, but seem to be blocked by the example rules previously, that force categories to be unstopped. This has been solved, somewhat inelegantly, by marking bracket-delimited phrases as unstopped, except in particular cases. When quote transposition occurs, for example, the stop value of the phrase inside the quotes should be the stop value of the whole quoted phrase. Also full-stop stopped phrases inside parentheses should make the whole bracket-delimited phrase full-stop stopped. Thus the rules for a quote-delimited noun phrase will look something like (4.20), in simplification.

(4.20) np[st –] ⇒ [punc q-open] np [punc q-close].
np[st @s] ⇒ [punc q-open] np[st @s] [punc q-close].

The introduction of the *stop* feature seems sufficient to cope with the punctuational phenomena introduced above. In order to incorporate the phenomena of interaction between punctuation and lexical expressions (e.g. preventing immediate nesting of similar delimited phrases), it is necessary to introduce a small number of additional features into the grammar.

If, for example, it is ensured that a dash-delimited noun phrase must have the feature-value pairing **[dash –]**, it can be further stipulated that any noun phrase that includes a dash-delimited phrase must receive the feature-value pairing **[dash +]**, so that the two cannot unify (4.21). This not only prevents immediate nesting (also for other punctuation marks, if other features are applied similarly) but also prevents attachment ambiguity (at least for the purposes of the parse) with the usage of certain punctuation marks (4.22). Thus in (4.23), *my friend — who sings* — would be marked as **[dash +]**, and hence would not be permitted as the dash interpolation attaching to the noun phrase *Luciano Pavarotti*. In (4.24), on the other hand, the dash interpolation to *Luciano Pavarotti* would only contain a constituent with **[comma +]**, which would be licensed by the dash interpolation rule (4.21).

(4.21) np[dash +, st @s] ⇒ np[st da] np[dash –, st –] [punc @s]

(4.22) ~ Luciano Pavarotti, my friend, an Italian, who sings, was on the plane.

(4.23) *Luciano Pavarotti — my friend — who sings — and also an Italian — was on the plane.

(4.24) Luciano Pavarotti — my friend, who sings, and also an Italian — was on the plane.

A similar rule to the above will prevent the nesting of colon-expansions, and therefore the relative scoping of colons and semicolons, as discussed in section 3.2 (examples (3.27) — (3.29)), can be incorporated into the grammar very easily too. If the features **colon** and **semi** represent the presence within a syntactic entity of a colon and semi-colon respectively, the rules can be designed as follows. The semicolon rule (4.25) should to apply regardless of the value of **colon** in its arguments, but ensuring that these arguments are **[semi –]**, (enabling phrases with or without a colon-expansion to be conjoined by a semi-colon, provided those phrases do not themselves contain semicolons). The colon rule (4.26), likewise, should apply only to arguments with **[colon –]**, irrespective of the value of **[semi]**. The mother categories of the rules will represent the content within the syntactic entity created by that rule of the relevant punctuation marks. Note that there are more versions of the colon rule, which deal with different constituents to either side of the colon, and also that, since the GDE does not permit the disjunction of feature values, the semicolon rule is merely an abbreviation of the multiple rules required in the grammar. Stop unification has been omitted for simplicity.

(4.25) s[semi +, colon (@a ∨ @b)] ⇒ s[semi –, colon @a, st sc] s[semi –, colon @b].

(4.26) s[semi (@s ∨ @t), colon +] ⇒ s[semi @s, colon –, st co] s[semi @t, colon –].

Thus sentences of the structure type exemplified in (4.27) will retain their (correct) ambiguous interpretations, with the semi-colon taking either wide (4.28) or narrow (4.29) scopings. Sentences with structures as (4.30) however, cannot receive ambiguous interpretations. The interpretations must be symmetrical (4.31) since otherwise the presence of positive **semi** or **colon** features would block the highest level rule-application (4.32).

(4.27) **e ; e : e** **e : e ; e**

(4.28) **(e ; (e : e))** **((e : e) ; e)**

(4.29) **((e ; e) : e)** **(e : (e ; e))**

(4.30) **e ; e : e ; e** **e : e ; e : e**

(4.31) **((e ; e) : (e ; e))** **((e : e) ; (e : e))**

(4.32) ***(e : (e ; (e : e)))**

In this manner, the inclusion of a few simple extra features in a normal grammar achieves an acceptable treatment of punctuational phenomena. However, since this work only represents the initial steps of providing a full and proper account of the role of punctuation, no claims are made for the theoretical validity or completeness of this approach! Thus this is only an *example* of a grammar of punctuation, and should not be regarded as the complete theory to be proposed later on in this thesis.

The Corpus

For the current investigation it was necessary to use a corpus sufficiently rich in punctuation to illustrate the possible advantages or disadvantages of utilising punctuation within the parsing process. Obviously a sentence which includes no punctuation will be equally difficult to parse with both punctuated and unpunctuated grammars. Similarly, for sentences including only one or two marks of punctuation, the use of punctuation is likely to be rather procedural, and hence not necessarily very revealing: a single punctuation mark in a sentence is almost certain to be a full-stop, which all grammars must cope with implicitly, since they are only fed a sentence at a time; a second punctuation mark is likely to be separating, something that unpunctuated grammars should be able to deal with in some unambiguous manner too.

Therefore the tagged Spoken English Corpus was chosen (Taylor and Knowles, 1988). This features some very long sentences, and includes rich and varied punctuation. Since the corpus has been punctuated manually, by several different people, some idiosyncrasy occurs in the punctuational style, but there is little punctuation which would be deemed inappropriate to the position it occurs in. Indeed, it could be argued that a degree of idiosyncrasy is helpful, since such idiosyncrasy is likely to be present in most punctuated texts produced without rigid stylistic prescription or editing.

A subset of 50 sentences was chosen from the whole corpus. Between them these sentences include material taken from news broadcasts, poetry readings, weather forecasts and programme reviews, so a wide variety of language is covered.

The lengths of the sentences varied from 3 words to 63 words, the average being 31 words; and the punctuational complexity of the sentences varied from one mark (just a full-stop) to 16 marks, the average being 4 punctuation marks. A sample tagged sentence is shown in (4.33), where **fs_FS** denotes a full-stop.

(4.33) Their_APP$ meeting_NN1 involves_VVZ a_AT1 kind_NN1 of_IO life_NN1 swap_NN1 fs_FS

The punctuated grammar, developed with this subset of the corpus, was used to parse the corpus subset, and then an unpunctuated version of the same grammar was used to parse an unpunctuated version of the same subset. The unpunctuated grammar could have been produced so that it just ignored the punctuation present in the original corpus, but for minor technical reasons this was not done. The reason that testing was performed on the training corpus was that, in the absence of a complete treatment of punctuation, the punctuational phenomena in the training corpus were the only ones the grammar could work with, and although they included almost all of the core phenomena mentioned, slightly different instances of the same phenomena could cause a parse failure. For reference, a small set of novel sentences were also parsed with the grammars, to determine their coverage outside the closed test. It should be stressed, however, that the current investigation is not about the coverage of the grammar, or indeed particularly about its design, but about the problem of whether a grammar that takes account of punctuation performs better than a grammar that is identical, except that it ignores punctuation. The definition of better performance here is fewer analyses (retaining the correct one), since the hits on correct analyses should be identical for the grammars, given one is an unpunctuated version of the other.

The unpunctuated version of the grammar was prepared by removing all the features relating to specifically punctuational phenomena, and also removing explicit mention of punctuation marks from the rules. This, of course, left behind certain rules that were functionally identical, and so duplicate rules were removed from the grammar. Similarly for rules which performed the same function at different levels in the grammar (e.g. attachment of prepositions to the end of a sentence with a comma in the punctuated grammar is also catered for in the unpunctuated grammar by rules allowing prepositions to be attached to the ends of noun and verb phrases).

Results

Results of parsing with the punctuated grammar were very good, yielding, on average, a surprisingly small number of parses. The number of parses ranged from 1 to 520, with an average of 38. This average is unrepresentatively high, however, since only 4 sentences had over 50 parses. These were, in general, those with high numbers of punctuation marks, all containing at least 5, as in (4.34). Ignoring the four smallest and four largest results then, the average number of parses is reduced to just 15. Examples (4.35) and (4.36) are more representative, both of the size and complexity of sentences in the corpus and also of parsing results. On examination, a great number of the ambiguities seem to be due to inaccuracies or overgenerality in the lexical tags assigned to words in the corpus. The word *more*, for example, is triple ambiguous as determiner, adjective and noun, irrespective of where it occurs in a sentence.

(4.34) The sunlit weeks between were full of maids: Sarah, with orange wig and horsy teeth, was so bad-tempered that she scarcely spoke; Maud was my hateful nurse who smelled of soap, and forced me to eat chewy bits of fish, thrusting me back to babyhood with threats of nappies, dummies, and the feeding bottle.


```
((More_DAR news_NN1
  (about_II
    (((the_AT Reverend_NNS1 Sun_NP1 Myung_NP1 Moon_NP1 cm_CM)
      (founder_NN1 (of_IO (the_AT Unification_NN1 church_NN1))) cm_CM)
     (who_PNQS
      ((’s_VBZ)
       ((currently_RR) in_II
        (jail_NN1 (for_IF (tax_NN1 evasion_NN1)))))) co_CO)))
 ((he_PPHS1)
  ((was_VBDZ awarded_VVN) (an_AT1 honorary_JJ degree_NN1)
   (last_MD week_NNT1
    (by_II
     (the_AT (Roman_JJ Catholic_JJ) University_NNL1
      (of_IO
       (la_AT Plata_NP1
        (in_II ((Buenos_NP1 Aires_NP1 cm_CM)
                (Argentina_NP1) fs_FS)))))))))
```

Figure 4.1: One possible punctuated parse of the sentence in (4.36)

(520 punctuated parses)

(4.35) The castaway in ’Desert Island Discs’, one hour late, is silly-ass actor Jeremy Lloyd, who’s also known as a scriptwriter. *(4 punctuated parses)*

(4.36) More news about the Reverend Sun Myung Moon, founder of the Unification Church, who’s currently in jail for tax evasion: he was awarded an honorary degree last week by the Roman Catholic University of la Plata in Buenos Aires, Argentina.
(18 punctuated parses)

Besides the ambiguity of corpus tags, a problem arose with words that had been completely mistagged. If these caused the parse to fail completely, the tag was changed in the development phase of the grammar, but even so, the number of complete mistags was rather small in the subcorpus used: around 10 words in the 50 sentences used.

The linguistic information that can be provided by the use of the punctuated grammar is shown in the representation of one of the possible parses of the sentence in example (4.36) in figure 4.1. The main features visible are the separation of the sentences into two main phrases, one to either side of the central colon, and the treatment of the two comma delimited noun phrases describing the subject of the first clause, *the Reverend Sun Myung Moon*. Whilst the correct unpunctuated parse of this sentence would also yield a similar structure as a possibility, there could be far greater ambiguity (more parses) and the unpunctuated parse will be unable to reflect such key punctuation-related linguistic facts as the colon indicating that the following phrase acts as an elaboration or explanation of the preceding one (Dale, 1991).

There did not seem to be any particular relationship between the length of a sentence and the number of parses, or between the quantity of punctuation marks and the number of

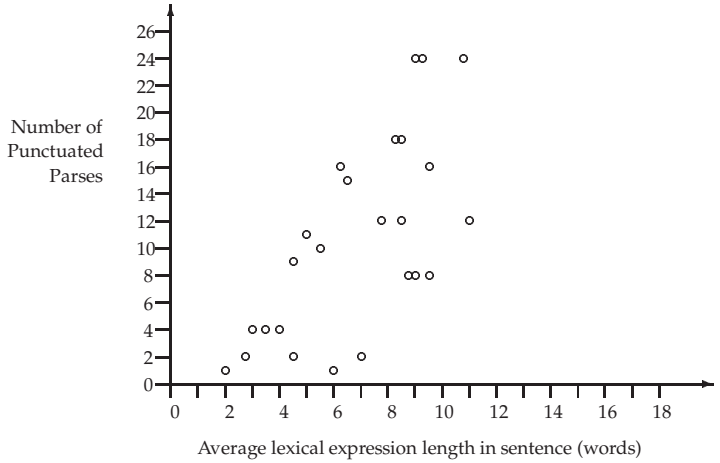


Figure 4.2: Plot of parse ambiguity against average lexical expression length with the punctuated corpus and grammar.

parses. Indeed several sentences with high numbers of punctuation marks, such as (4.35) which contains 6, and similarly several long sentences, such as (4.37), had low ambiguity. However if the two quantities are related, so that we obtain the average length in a sentence of the lexical expressions between punctuation marks, a better relationship emerges. Figure 4.2 shows a plot of the average expression length against parse ambiguity, for the central 25 sentences of the corpus subset, in terms of their ambiguity (so that very low or very high numbers of parses have been omitted). Whilst the plot is by no means an indication of any formal relationship, it does show the upward trend of parse ambiguity with length of lexical expressions. And in particular, it shows that there are no sentences that have short average lexical expression lengths but high ambiguity. It is interesting to compare this to the general results when parsing sentences conventionally (without punctuation), where greater sentence lengths tend to imply greater parse ambiguity, since it shows that the use of punctuation, in effect, breaks down the sentence into more manageable chunks.

(4.37) And in London expect long delays in Chiswick, where the westbound elevated section of the M4 is closed for most of today, and only a single eastbound lane is open. *(2 punctuated parses)*

Initial attempts at parsing the corpus subset using the unpunctuated version of the grammar were unsuccessful on even the most powerful machine available. This was due to the failure of the machine to calculate the number of parses in the parse forest produced by the chart parser without enumerating them, which would have taken too much memory. A special

section of code written for the GDE (grateful thanks are due to John Carroll for supplying this piece of code) to estimate the number of individual parses represented by the packed parse-forest showed that for all but the most basically punctuated sentences, the number of parses was ridiculously huge. The figure for the sentence in (4.36) was in excess of 6.3×10^{12} parses! Even though this estimate is an upper bound, since effects of feature value percolation during unpacking are ignored by the code, it has been fairly accurate with most grammars in the past and still indicates that rather too many parses are being produced! Not all sentences produced such a massive number of parses: the sentence in (4.38) yielded only 192 parses with the unpunctuated grammar which was by far the smallest number of unpunctuated parses. Most sentences that managed to pass the estimation process produced between 10^6 and 10^9 parses.

(4.38) Protestants, however, are a tiny minority in Argentina, and the delegation won't be including a Roman Catholic. *(9 punctuated parses)*

On closer examination of the grammar and the corpus, it is possible to understand how and why this has happened. In the absence of punctuation marks, ambiguity arises where syntactic constituents that usually require some punctuation to be present become confused with similar constituents that do not. For example, the punctuated grammar needs to allow for complex noun phrases that include comma-delimited noun phrases (non-restrictive relative clauses), undelimited noun phrases (restrictive relative clauses) and also compound noun phrases (4.39). These are relatively easy to mark and recognise when the punctuation is available, but without punctuational clues and with the under-specific tagging system used by this grammar these sentences all appear identical without punctuation. Thus the unpunctuated grammar contains three more-or-less identical rules to process complex noun phrases, which all apply in every circumstance given in (4.39). Hence the multiple rules to process the complex NPs and also other syntactic categories must be removed in the unpunctuated grammar, since they will all be acting in parallel.

(4.39) My friend, the Italian, is well.
My friend the Italian is well.
The Rank Hovis McDougall Company today announced…

Therefore the unpunctuated grammar was further trimmed, to such an extent that parses no longer accurately reflected the linguistic structure of the sentences, since, for example, comma delimited noun phrases and compound nouns became indistinguishable. Some manual preparation of the sentences was also carried out to prevent the re-occurrance of these simple but costly misparses.

The results of the parse now became much more tractable. For basic sentences, as predicted, there was little difference in the performance of punctuated and unpunctuated grammars. Results were within an order of magnitude, showing that no significant advantage was gained through the use of punctuation. The sentences in (4.40), (4.41) and (4.42) received 1, 11 and 84 parses respectively with the unpunctuated grammar.

(4.40) Well, just recently, a day conference on miracles was convened by the research scientists, Christian Fellowship. *(4 punctuated parses)*

(4.41) The assembly will also be discussing the UK immigration laws, Hong Kong, teenagers in the church, and of course, church unity schemes. *(2 punctuated parses)*

(4.42) Next week a delegation of nine Protestant ministers from Argentina visits the Autumn assembly of the British Council of Churches; it's meant as a symbol of reconciliation between Christians, following the Falklands War. *(12 punctuated parses)*

For the most complex sentences, however, the number of parses with the unpunctuated grammar was typically around two orders of magnitude greater than with the punctuated grammar, and even higher with certain sentences. For examples, the sentence in (4.43) had 12,096 unpunctuated parses.

(4.43) They want to know whether, for instance, in a scientific age, Christians can really believe in the story of the feeding of the five thousand as described, or was the miracle that those in the crowd with food shared it with those who had none?
(24 punctuated parses)

Parsing a set of ten previously unseen punctuationally complex sentences with the punctuated grammar resulted in seven of the ten being unparsable. The other three parsed successfully, with the number of parses falling within the range of the results of the first part of the investigation. The parse failures, on examination, were due to novel punctuational constructions occurring in the sentences which the grammar had not been designed to handle. Parsing the unseen sentences with the unpunctuated grammar resulted in one parse failure, with the results for the other 9 sentences reflecting the previous results for complex sentences.

Summary

An explanation that should be made concerns the ultimate goals of the study of punctuation. The aim is not a genre-specific all-inclusive grammar of punctuation, but rather a general set of ideas that will indicate the function and meaning of a given mark of punctuation in any position that it occurs in. The current study forms a justification for such work. Since the goal is not genre-specific, the facts that the SEC is transcribed speech, and that it is idiosyncratic in punctuation use, according to the whims and linguistic intuitions of the transcribers, do not hinder the investigation in any way.

The current investigation seems to support the original premise — that inclusion and use of punctuational phenomena within natural language syntax can assist the general aims of natural language processing.

We have seen that for the simplest sentences, use of punctuation gives us little or no advantage over the more simple grammar, but, conversely, does no harm and can reflect the actual linguistic construction a little more accurately.

For the longer sentences of real language, however, a grammar which makes use of punctuation massively outperforms an otherwise similar grammar that ignores it. We also see that

unlike unpunctuated grammars and texts, where there is some relation between sentence length and parse ambiguity, the ambiguity for punctuated sentences is better related to the lengths of the sequences of lexical expressions that occur between the punctuation marks. It is difficult to see how any grammar that takes no notice of punctuation could ever become successful at analysing such long, punctuated sentences unless some huge amount of semantic and pragmatic knowledge is used to disambiguate the analysis.

There are, however, several limitations to the current approach. In the grammar, there are several inelegant processes, such as in (4.20), which just served as working solutions to problems encountered. It is more than likely that deeper investigation into punctuational theory will yield better, more elegant ways of solving such problems. Similarly, adding punctuation to a grammar increases its complexity, which can compensate for the removal of other ambiguity. It would be helpful if the increase in complexity could be minimised.

As was shown by the attempt at parsing the novel sentences though, knowledge of the role of punctuation is still severely limited. The grammar only performed reliably on those punctuational phenomena it had been designed with. Unexpected constructs caused it to fail totally. Following the recognition that punctuation can play a crucial role in natural language syntax, therefore, the way is clear for the development of a full theory of the function of the punctuation system. This development will begin in the following chapter with an exploration of the variety and usage of the physical punctuation marks themselves, before progressing in subsequent chapters onto the more detailed aspects of syntactic and semantic functionality.

five Exploring the Variety and Use of Punctuation

They were in Derek's bedroom.
Ray came back in.
— I was thinkin' there, he said. — I think maybe we should have an exclamation mark, yeh know, after the second And in the name.
— Wha'?
— It'd be And And exclamation mark, righ', And. It'd look deadly on the posters.
— Outspan said nothing while he imagined it.
— What's an explanation mark? said Derek.
— Yeh know, said Ray.
He drew a big one in the air.
— Oh yeah, said Derek. — An' where d'yeh want to put it again?
— And And
He drew another one
— And
— Is it not supposed to go at the end?
*— It should go up his a***, said Outspan, picking away at the sticker.*

(Doyle, 1988)

Only now that it has become clear that punctuation is useful to Natural Language Processing, and also that a justification for the inclusion of punctuation treatments in natural language processing systems has been provided, is the path clear to begin the construction of a punctuation theory. An ideal place to start, it seems, is in the examination of punctuation itself. It will be crucial to further investigations to know what all the likely marks of punctuation are (and even to know what the unlikely ones might look like!) and one of the fundamentals of a punctuation theory must be the interaction of individual marks of punctuation. Indeed, in the work of Nunberg (1990), this interaction is a basis of an important portion of his whole analysis, with principles such as point absorption and quote transposition. To try to discover the identity and interaction of punctuation, then, a small pilot study was carried

out, and the results were used to guide and develop a later, larger and more comprehensive study.

5.1 Pilot study

Several corpora were processed to have their punctuation patterns extracted. The corpora in question were sections from two consecutive years (1990 and 1991) of the Guardian Newspaper (approximately 12 million words), the Leverhulme corpus (356,000 words) and the familiar Wall Street Journal corpus (184,000 words). These were automatically processed to yield the punctuation patterns in the sentences, so that all punctuation marks were reported as they occurred in the text, but sequences of words occurring between punctuation marks were just reported as a single token. Thus this paragraph might have been reported as the pattern in (5.1).

(5.1) **[e . e (e) , e (e) e (e) . e , e , e . e (e) .]**

The punctuation sequence so obtained was split into patterns corresponding to the separate sentences of the text. The procedure for achieving this was not simply a segment-upon-reaching-dot procedure, but also examined the environment of any dot found. Dots in the middle of text (decimal points, for example) did not cause segmentation, and sentence-final dots occurring inside quotes or brackets similarly did not cause segmentation. A dot occurring just before a final quotation mark, reflecting the familiar process of quote transposition (Nunberg, 1990), did correctly trigger a sentence segment, though. A similar treatment as for a full-stop was given to other sentence boundary markers such as question marks and exclamation marks. Thus the procedure for recognising sentence boundaries was roughly as illustrated in figure 5.1.

Whilst the segmentation processes described above were not suitable to recognise 100% of the sentence boundaries, failing for such cases as abbreviations, they correctly recognise the great majority of sentence boundaries. The problem of correctly identifying a sentence boundary 100% of the time, whilst unquestionably important and interesting, was too great a task for the current investigation. Some research describing this problem can be found in (Palmer and Hearst, 1994).

Identical sentential patterns emerging from the above processing were grouped together, yielding frequency information for their occurrence. In addition, these patterns, with frequency information, were further processed to give the frequency of occurrence of each individual punctuation mark in a given corpus.

Punctuation statistics

The following tables, 5.1 – 5.4, show the gross numerical results for the frequencit of selected punctuation marks in all of the corpora analysed. For comparison, the results of a similar study by Meyer (1987) on the Brown corpus of American English are also shown in table 5.5. A distinction has been made between the point marks (commas, full-stops, question and exclamation marks, colons, semi-colons and dashes) which can occur singly, and the bracket marks

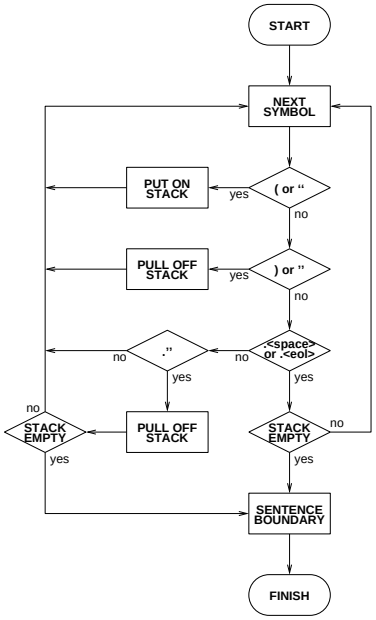


Figure 5.1: Flowchart to illustrate sentence division.

(parentheses and quotation marks) which always occur in matched pairs. To a certain extent it could also be argued that the point punctuation symbols have a more linguistic usage, whereas the bracket symbols are more orthographic/stylistic. In any case, results are presented in the tables 5.1 – 5.5 below in terms of percentage of point symbols alone, and also percentage of all punctuation marks in the corpus, including bracket symbols. Since bracket symbols always occur in matched pairs, their entries in the tables give the total number of such symbols, of which half will be opening symbols, and half closing symbols.

There are several quirks present in the processing of these corpora and the presentation of the results. Due to storage restrictions in the initial stages of analysis, for the 1990 section of the Guardian corpus, only those punctuation patterns with a frequency of occurrence above six times in the corpus were stored, analysed and presented in these results. The quotation marks present in the Wall Street Journal corpus included some very bad mismatched symbols, causing a problem to the automatic analysis stage. Therefore, for this corpus, quotation marks are not included in the reported results. Quotation marks are similarly not reported for the Leverhulme corpus, since a totally different representation was used for the marks. It was very difficult to try to disambiguate opening and closing quotation marks in this representation, once again causing problems for the automatic analysis. Furthermore, it should be made clear

Mark	Total	Point %-age	Total %-age
Comma	204,011	49.7	43.2
Full-stop	189,681	46.2	40.2
Quotation-marks	50,282		10.7
Parentheses	10,740		2.3
Colon	8,112	2.0	1.7
Question-mark	4,442	1.1	0.9
Semi-colon	3,542	0.9	0.8
Dash	624	0.2	0.1
Exclamation-mark	341	0.1	0.1
Point Total	410,753	100.2	
Total	471,775		100.0

Table 5.1: Punctuation results for segments of The Guardian, 1990

Mark	Total	Point %-age	Total %-age
Comma	297,834	50.5	42.7
Full-stop	255,274	43.3	36.6
Quotation-marks	77,264		11.1
Parentheses	31,020		4.4
Colon	15,026	2.5	2.2
Semi-colon	10,827	1.8	1.6
Question-mark	6,813	1.2	1.0
Dash	2,702	0.5	0.4
Exclamation-mark	944	0.2	0.1
Point Total	589,420	100.0	
Total	697,704		100.1

Table 5.2: Punctuation results for segments of The Guardian, 1991

Mark	Total	Point %-age	Total %-age
Comma	8,250	46.4	46.2
Full-stop	7,683	43.2	43.0
Semi-colon	969	5.4	5.4
Dash	444	2.5	2.5
Question-mark	188	1.1	1.1
Colon	156	0.9	0.9
Exclamation-mark	94	0.5	0.5
Parentheses	82		0.5
Point Total	17,784	100.0	
Total	17,866		100.1

Table 5.3: Punctuation results for the Wall Street Journal

Mark	Total	Point %-age	Total %-age
Full-stop	15,453	56.1	51.9
Comma	9,636	35.0	32.4
Parentheses	2,230		7.5
Dash	1,161	4.2	3.9
Semi-colon	385	1.4	1.3
Question-mark	352	1.3	1.2
Colon	343	1.2	1.2
Exclamation-mark	199	0.7	0.7
Point Total	27,529	99.9	
Total	29,759		100.1

Table 5.4: Punctuation results for the Leverhulme Corpus

Mark	Total	Point %-age	Total %-age
Comma	4,054	47.7	46.8
Full-stop	3,897	45.9	45.0
Dash	189	2.2	2.2
Semi-colon	167	2.0	1.9
Parentheses	165		1.9
Question-mark	84	1.0	1.0
Colon	78	0.9	0.9
Exclamation-mark	25	0.3	0.3
Point Total	8,494	100.0	
Total	8,659		100.0

Table 5.5: Punctuation results for the Brown Corpus

that a single – character was considered to represent a dash, but that the extraction software checked to ensure the symbol was surrounded by spacing to avoid ambiguity with hyphens.

Several interesting points emerge from examination of tables 5.1 – 5.5:

- The comma and full-stop occur with greater frequency than any other mark. When only the point punctuation is considered, the distance widens since the next highest point mark in any corpus accounts for less than 5.4% of the total.
- Some clear differences between genre emerge: the relatively balanced and formal corpora, the Brown corpus and, to a certain extent, the Guardian and WSJ (since these journalistic corpora contain a fairly wide cross-section of language, including news, prose and speech), which involve more sophisticated, structured language, contain more commas than full-stops so that every sentence will contain, on average, at least one comma and hence is likely to have a complex phrasal structure. In the Leverhulme corpus, however, which consists of essays by students in secondary education, the situation is reversed. There are more full-stops than commas, indicating a simpler use

of language.

- In the comparison of the two years of the Guardian, some significant differences in results emerge. This is likely to stem from truncation of the frequency results in the 1990 corpus, since the results would be expected to be broadly similar due to formal editing and use of prescriptive style-guides. The reported increase in the frequency of complex sentences in 1991 (suggested by a higher percentage of commas, and a lower percentage of full-stops) means, unsurprisingly, that the low frequency sentences are more complex than the high-frequency ones. This is borne out by the increase in frequency of all the minor point marks. Particularly significant are the increases in frequency of the colon (a 25% improvement from 2.0% to 2.5%) and the semi-colon (a 100% improvement from 0.9% to 1.8%). This probably reflects the presence in the low-frequency sentences of a great many long, complex, lists. Every day, for instance, a long list of birthdays is given, in which each item has the structure *[name , occupation , age]* and items are semi-colon separated.
- In the only corpus not to have been produced with the use of a style guide, the Leverhulme, it is interesting to note that the difference between frequencies of full-stops and commas is quite large. There are roughly two-thirds more full-stops than commas, indicating a high number of simple, declarative sentences. The incidence of the dash is also very high, at 4.2%, compared with the next highest incidence in other corpora of 2.9%. This suggests that when people are writing formally, fewer dashes are used than when writing freely. This is particularly noticeable in personal letter-writing, where dashes are often used to indicate a mental pause in writing, or to separate items in a 'stream of consciousness'. Another significant difference is the frequency of the exclamation mark, which occurs almost twice as frequently (0.7%) in the Leverhulme corpus as in any other. This reflects the popular fact that people overuse this mark, and use it wrongly, or idiosyncratically, unless taught otherwise. Hearteningly, the relatively high incidence of the semi-colon, suggests that the hypothetical architypal retired schoolmaster in the quote in chapter 3 may be incorrect, although there is, of course, no guarantee that the semi-colons are used appropriately.
- As a follow-up to the previous observations, stress-markers (question and exclamation marks) are used infrequently in comparison to full-stops. Of these markers, however, question marks are far more common than exclamation marks. As observed above, stress markers are significantly more common in non-style-guided material, especially the exclamation mark.
- As a by-product of the analysis techniques, ellipses *[...]* were not reported. Their use is relatively infrequent though, judging by manual examination of the corpora, and they are far less frequent in style-guided corpora than in free ones. Their use, to indicate the possibility of a sentence, theme or 'stream-of-consciousness' continuation that has not been written explicitly, is a very informal one. In formal writing it is possible to indicate something similar through textual means, through use of phrases such as *etcetera* or *and*

others, and the requisite possibility of continuation in the case of a conceptual discontinuity can be inferred by the reader without any such device at all.

- The trouble with the Wall Street Journal corpus and the Brown corpus is that they are relatively small, with only 8,000 and 4,200 sentences respectively. Therefore, the results from these corpora are likely to be slightly less valid and conclusive than results from the larger corpora: the Leverhulme at 16,000 sentences, and the Guardian at 260,000 sentences per fragment per year.

- Interestingly, despite the above observations of the relative simplicity of sentences in the Leverhulme corpus, the average sentence length remains broadly similar across the corpora: 23 for WSJ and Guardian, and 22 for the Leverhulme. This means either that the Leverhulme corpus contains more ambiguity in its sentences and far less structure, or that the corpora using more punctuation are over-using it. Since over-use of punctuation has the potential to be as confusing as under-use, it seems far more likely that the former conclusion is the correct one, since high-profile formal publications are unlikely to deliberately set out to confuse readers.

- The problem of the corpora produced with style-guides is that by definition, their use of punctuation reflects the style guide rather than the language. Of course, it is not the case that style guides do not reflect the language, just that their use restrains and restricts it somewhat. Hence from the point of view of analysing punctuation patterns occurring in real text, the Leverhulme corpus, with all its 'incorrect' uses of punctuation, is likely to be of more real use to the ongoing investigation.

Sentence complexity

It is also interesting to study the individual patterns of punctuation use that emerge from the analysis of the corpora. Considering sentences that contain only commas and the final full-stop, there is a pattern in all corpora of decreasing frequency with increasing sentence complexity (number of commas) shown in table 5.6 and represented graphically in figure 5.2. These results are only to be expected really, since it seems clear that in any given text the more complex a sentence structure, the less frequently such a structure will appear, but some interesting observations can still be made:

- Again there is a great similarity within genre. The results for the two years of the Guardian are almost identical, and results for the Wall Street Journal are at least comparable, with rather more single-clause sentences, and fewer sentences with more than two clauses. The results for the Leverhulme corpus, however, show a confirmation of the observations made above: there are far more basic sentences, without any commas, and far fewer sentences containing commas. In fact, there are almost double the number of single-clause sentences in this corpus as double-clause sentences. Furthermore, the Leverhulme corpus, which should be large enough, shows none of the very complex sentences found in the Guardian, and which are hinted at in the much smaller Wall Street

No. of commas in sentence		Guardian 1990		Guardian 1991		Wall Street Journal		Leverhulme	
0	(e.)	64,599	44%	80,606	43%	2,713	56%	3,940	63%
1	(e,e.)	38,882	26%	48,273	26%	1,237	25%	1,440	23%
2	(e,e,e.)	27,282	18%	35,123	19%	512	10%	628	10%
3	(e,e,e,e.)	10,759	7%	13,747	7%	221	5%	188	3%
4	(e,e,e,e,e.)	4,048	3%	5,274	3%	107	2%	53	1%
5	(e,e,e,e,e,e.)	1,541	1%	2,012	1%	41	1%	22	
6	(e,e,e,e,e,e,e.)	569		750		27	1%	6	
7	(e,e,e,e,e,e,e,e.)	234		306		12		4	
8	(e,e,e,e,e,e,e,e,e.)	98		127		3			
9	(e,e,e,e,e,e,e,e,e,e.)	44		53		7			
10	(e,e,e,e,e,e,e,e,e,e,e.)	21		25		2			
Total		148,077		186,296		4,882		6,281	

Table 5.6: Frequencies of increasingly complex sentences.

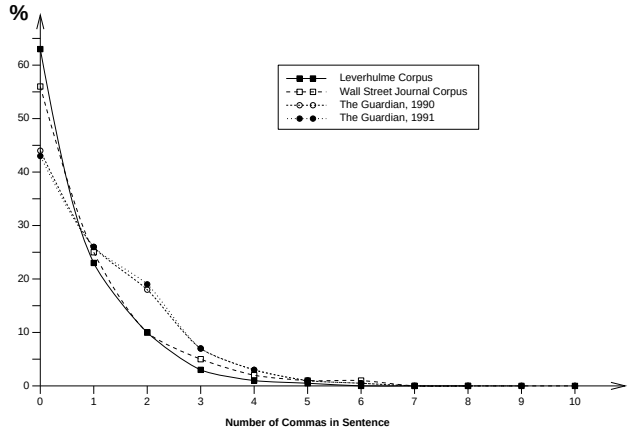


Figure 5.2: Graph of the frequency of complex sentences.

Journal corpus. Thus whilst the WSJ is very small, it still includes some examples of the highest complexity sentences. Since the frequency of occurrence should be related to the length of the corpus, we would expect the longer Leverhulme corpus to include at least some of the most complex structures. However, since it includes no sentences above a comma inclusion of seven, it must be deduced that the language contained in this corpus is fundamentally different to that in the other corpora.

- As the totals for each column in table (5.6) show, the sentences shown here do not represent the whole corpus. While the inclusion of sentences containing colons or semi-colons would be meaningless since the sentences would not have the same basic clausal pattern, other sentences which fit the patterns, i.e. a basic sentence with *n* clauses in it, have been omitted. Examples of such sentences are those in which a clause might contain internal structure, as illustrated in example (5.2) which although a three clause sentence, would not have been included in the results shown in table 5.6 by the automatic extraction process. However, it is presumably the case that such features occur uniformly across the corpus, and hence whilst the overall frequency results will be affected, the relative frequencies will not.

  (5.2) **[e , e (e) e , e .]**

- Similar results to those in table 5.6 are seen with other sentence-final marks, such as question marks. The same basic patterns emerge, with more complex sentences occurring less frequently. Again, the relative rarity of the exclamation mark with respect to the question mark is reflected in these results, and even from the analysis of around 210,000 sentences, none of the basic clausal sentences containing exclamation marks had more clauses than three. Tables 5.7 are from The Guardian, 1990.

- The more complex punctuational sentences, such as those including colons, reflect the same pattern. The more complex the basic clausal sentence patterns to either side of the main point punctuation mark, the less frequent the incidence of the punctuation pattern in the corpus. Results in tables 5.8 and 5.9 are again taken from The Guardian, 1990.

- As with comma-separated clauses, similar patterns can also be seen with semi-colon separated clauses. Table 5.10 is taken from The Guardian, 1990.

Punctuation 'correctness'

The question now arises of the notional correctness of the punctuational patterns produced from these corpora. Although the aim of this account is not to provide a list of strictly correct and incorrect usages of punctuation, it is useful to compare the punctuation usages observed in real text with the prescriptive accounts of punctuation present in the style guides, and even with the limited degree of prescription present in the theory of (Nunberg, 1990). Since this latter is more easily codifiable, the patterns emerging in the results will be compared with the notions of correctness introduced by Nunberg. Excluding quotation marks, Nunberg's notions of punctuation patterns are those that will be accepted by the following context-free grammar

Pattern	Frequency
e?	2,520
e,e?	606
e,e,e?	357
e,e,e,e?	85
e,e,e,e,e?	38
e,e,e,e,e,e?	12

Pattern	Frequency
e!	161
e,e!	58
e,e,e!	23

Table 5.7: Frequencies of increasingly complex stress-marked sentences

Pattern	Frequency
e.	64,599
e:e.	1,660
e:e,e.	639
e:e,e,e.	385
e:e,e,e,e.	189
e:e,e,e,e,e.	80
e:e,e,e,e,e,e.	32
e:e,e,e,e,e,e,e.	16
e:e,e,e,e,e,e,e,e.	10

Pattern	Frequency
e.	64,599
e:e.	1,660
e,e:e.	335
e,e,e:e.	188
e,e,e,e:e.	45
e,e,e,e,e:e.	12

Table 5.8: Frequencies of increasingly complex colon-sentences.

Pattern	Frequency
e.	64,599
e;e.	885
e;e,e.	299
e;e,e,e.	171
e;e,e,e,e.	41
e;e,e,e,e,e.	18
e;e,e,e,e,e,e.	6
e;e,e,e,e,e,e,e.	6

Pattern	Frequency
e.	64,599
e;e.	885
e,e;e.	262
e,e,e;e.	134
e,e,e,e;e.	47
e,e,e,e,e;e.	8

Table 5.9: Frequencies of increasingly complex semicolon-sentences.

Pattern	Frequency
e.	64,599
e;e.	885
e;e;e.	73
e;e;e;e.	26
e;e;e;e;e.	21
e;e;e;e;e;e.	13

Table 5.10: Frequencies of increasingly complex semi-colon separated sentences.

(note that this grammar may give ambiguous and incorrect analyses, and hence is merely an indicator of validity rather than structure). The reason that quotation marks are excluded is that due to the phenomenon of quote transposition, representation would be very difficult (we would need to maintain a stack of matching punctuation structures entered and exited, which perhaps will have some bearing on the debate of the context-freeness of natural languages), and in practice quotation marks can be inserted anywhere in a sentence as long as they are matched, nested and do not interfere with parenthesis nesting. Thus the pattern in (5.3) would not be permitted.

(5.3) *[e (e “ e) e . ”]

The treatment of parentheses in this grammar is also rather simplistic, since it does not cater for the case where multiple sentences can occur within the parentheses. However, the main purpose of this grammar is to determine the validity of point punctuation patterns, rather than any bracket patterns. According with Nunberg's principles, this grammar forbids multiply-nested colon-expansions, but does permit a further colon-expansion within a dash interpolation or parenthetical. Also point absorption of final dashes is catered for. It should be stressed that this grammar is in no way linguistic, and does not generate any linguistically-relevant structure. It merely ensures that the punctuation marks in the text are used appropriately.

```
S  -->  T .  |  T ?  |  T !

T  -->  C  |  T ; C

C  -->  D  |  D': D

D' -->  P  |  P - T - P

D  -->  P  |  P - T - P  |  P - T

P  -->  L  |  P , L

L  -->  e  |  e ( T ) e  |  e ( T )
```

Using this simple grammar, the vast majority of sentence patterns are 'legal' by Nunberg's definition, but there are a few which are not:

- **Multiple colon-expansions** — Patterns of the sort shown in (5.4) were found in several corpora, from the Guardian to Leverhulme. Under Nunberg's definition, this pattern is incorrect. However, the sentences found in the corpora seem perfectly understandable, and correct in usage (5.5) and (5.6). The inapplicability of Nunberg's restriction on the colon-expansion is similarly pointed out in (Sampson, 1992), with felicitous examples, and so it seems that Nunberg's theory veers too far into prescription. Of course, sentences that do not nest colon-expansions are less ambiguous and look neater, so the principle

should be applied in generation, or formal editing. However, since instances of the pattern occur in analysis, it should be catered for.

(5.4) **[e : e : e .]**

(5.5) Therefore the eye counters this by having a built in tremor which ensures that the image passes from one set of adjacent cells to the next and back again in rapid succession so that no single group becomes depleted: in fact experiments which fix an immobile image on the eye show that subjects quickly become 'blind' to the stimulus: this is of course a significant difference in itself to the camera where such practices would be anathema due to the blurring this would cause.

(5.6) Here are some of the main arguments that have been put forward by each camp: The case for war: The argument in favour of going to war to remove Iraq from Kuwait is quite simple: Saddam Hussein has used naked aggression against a small and defenceless country.

- **Symbol Repetition** — such punctuation usage, illustrated in (5.7) occurs almost exclusively in the Leverhulme corpus, reflecting the untraditional nature of such use. It is a practice which is conventionally discouraged by style guides and teachers, but can have a pragmatic interpretation, as discussed in (Parkes, 1992) and mentioned in section 10.6, and so should not be ignored in analysis. For simplicity, sequences of such symbols could be treated as a single symbol, although at a more semantic or pragmatic level there may sometimes be arguments for marking such sentences with a higher stress than would be attributable to just a single stress marker. Examples of such symbol repetition are shown in (5.8–5.11). Example (5.8) is the only case of such symbol repetition in a formal corpus (The Guardian), and indicates an omission in editing rather than punctuation. The slot marked by *???????* should have been filled with a number before publication. Example (5.9) illustrates a similar unfilled slot case.

(5.7) **[e??? e!! e??!! e——e.]**

(5.8) However, BSkyB does say that sales of satellite equipment are, despite the recession, still increasing and currently stand at ???????

(5.9) A second factor worth considering is whether the order of presentation of a p—— would affect its impact.

(5.10) The question now was, which other blue dress??!!

(5.11) Over half a million young people simultaneously looking for the meaning of life !!

- **Mistaken Insertion** — There are some instances of inexplicable mis-punctuation, mainly in the Leverhulme corpus, which are almost certainly artifacts of the transcription into machine-readable form rather than of the original text. Thus in examples (5.12), a faulty keystroke has placed a full-stop adjacent to a comma.

(5.12) I will begin at the most obvious,. though not necessarily the most simple level of attempting to outline a definition of modernity.

- **Novel Constructions** — There are some instances of strange compound punctuation marks in the Wall Street Journal (5.13) which Nunberg's theory would not sanction, but which are obviously licensed by their style guide. In the first sentence (5.14), the novel punctuation could be replaced by either a normal comma or a normal dash, and in the second (5.15) novel punctuation could be replaced by just a colon. The novel punctuation in the latter case, of course, is quite a commonly used form despite not being mentioned in Nunberg's theory. Thus these are stylistic variants. It is interesting to note that these stylistic alteration show that Nunberg's principles of punctuation interaction are also stylistically determined, in that the dash in (5.14) does not absorb the adjacent comma.

(5.13) **[e,–e. e:–e.]**

(5.14) Claude rose and dressed, — a simple operation which took very little time.

(5.15) At first the gov't tried to 'play it down' for various reasons:– it would detract from the war effort, it had too many socialist undercurrents, it would be too expensive to implement, but public outcry was so great that this did not work.

- **Critical Punctuation** — The curious punctuation pattern found in (5.16) seems mainly to indicate an omission of uncertainty. The pattern is also occasionally used next to an uncertain fact (e.g. There are 9 (?) planets in the solar system.). Note that in the example below, the opening parenthetical is an illegal construct according to Nunberg. In fact, on examination of the context, the parenthetical should have been attached to the end of the previous sentence, and the word *but* capitalised.

(5.16) (One chromosome from mother: one from father) but what actual characteristics the child develops, depends on whether the genes are coded for the same characteristicie blue eyes or for different ones (one for blue eyes one for brown) in this case, where the genes are heter(?) one gene is usually dominant and one recessive.

In addition it is clear from examination of the results that there are sentences in all corpora that contain quote transposition *["e."]* and also those that ignore it *["e".]*. Hence Nunberg's claim that transposition is an American-English phenomenon which does not occur in British-English is not borne out — a point of view that also emerges from (Sampson, 1992).

The results from this section of the investigation confirm not only that the most important punctuation marks are the full-stop and comma, and therefore that development of a correct treatment of these will be of the greatest benefit to the field of language analysis, but also that the strictly prescriptive style guides, and also the linguistic treatments of Nunberg (1990), whilst suitable for production of text, are too prescriptive for the analysis of raw text since punctuation patterns that do not occur or are actively disapproved of, do occur in texts. Thus any treatment should have the capacity to assign at least some meaning to the 'incorrect'

punctuation patterns, otherwise systems will be of little use. This is more true in the field of punctuation than the related problems in the field of syntax, since there is less regularity in the use of the punctuation system, and a great deal of idiosyncratic usage occurs.

5.2 Full study

The preliminary study has shown some very interesting and promising results, but there were some problems associated with it:

- Insufficient data — only three corpora were analysed (counting the two years of The Guardian together) and there was a huge discrepancy in their size, The Guardian being about twelve times larger than the other two together.
- Punctuation preconceptions — the original analyser was designed with a fairly rigid notion of punctuation embodied. Bracketing symbols and quotes had to match, otherwise the sentence was not terminated, and only a small sub-domain of non-lexical orthography was reported.
- Punctuation orientation — related to the previous point, the results tended to be oriented towards specific punctuation marks, or at least those which had been considered in the design of the system. Unfortunately, a consequence of this was to completely omit all complex punctuation marks (those consisting of two or more symbols).
- Corpus-specific orientation — the system was designed with The Guardian, and hence its notions of several marks, for example quotes, are based on those of The Guardian, hence the problems with quotation marks in the other corpora.
- Structure-blindness — the system did not take any account of structural cues, such as paragraph-breaks (blank lines) or new files. These help disambiguate unmatched punctuation.

The system was therefore redesigned to take account of any non-alphanumeric character in the text it was examining. All forms of parenthesis were treated properly by the bracket-matching system, as were all forms of quotation mark. Since there are potentially four different ways of quoting

'single quotes'
"double quotes"
`'single keyboard quotes'`
`"double keyboard quotes"`

and some of the quotation marks conflict across the different types, one of the key cues the system took was whether the marks occurred immediately before or after a word. Because of this, and the consideration of all four methods rather than just the second, the system also reports word-final apostrophes. To ensure that these did not interfere with the matching

Corpus	Size (words)	Size (sentences)	Mean sentence length
The Guardian [1990]	23,963,515	961,604	24.9
The Guardian [1991]	21,638,956	879,438	24.6
Leverhulme	355,594	15,547	22.8
The King James Bible	820,731	30,021	27.3
The Quran	168,663	14,312	11.8
Shakespeare (complete works)	939,193	63,692	14.7
Spoken English Corpus	54,614	2,930	18.6
IPPE Philosophy	518,138	28,945	17.9
Project Gutenberg	13,747,367	649,069	21.2
Usenet	22,779,757	1,658,707	13.7
Total	84,986,528	4,304,265	19.7

Table 5.11: The Corpora used for punctuation extraction.

process, the system was designed not to protest at any unmatched final quote marks that appeared in the text.

As before, the redesigned system detects sentence ends at full-stops, so long as the closure list is empty. The closure list has a symbol placed upon it for every opening matched punctuation mark that is detected, and these symbols are removed when the corresponding closing mark is reached. Further to this, full-stops immediately followed by punctuation marks that cause the closure list to become empty are also treated as sentence-final (although of course the further punctuation marks are also reported as part of that sentence). Therefore the familiar process of quote transposition (Nunberg, 1990) has been accounted for. Unlike the system used in the pilot study, if a closing matched mark is omitted, or if a full-stop is omitted, the redesigned system does not treat the rest of the corpus as a part of the same sentence. If a blank line is encountered (which is one of the most common ways to represent paragraphing), or an end-of-file marker, then a sentence is regarded as terminated regardless of what is on the closure list.

Furthermore, it should be stressed that only word-external punctuation is examined — that is punctuation that is bounded at one or both sides with whitespace. Therefore decimal points, hyphens and certain apostrophes will be ignored, since they are word-internal. Since they are marks of sub-lexical punctuation rather than the inter-lexical ones that are being investigated in this study, though, this is not a problem.

The rest of the processing was as in the pilot study. Sentential punctuation patterns were counted, yielding the frequency of occurrence for each pattern, and these frequencies were further processed to yield data for individual punctuation symbols.

A far greater range and quantity of source text was used in the corpora for this study, and around 85,000,000 words were processed. The corpora that were chosen for processing in this study are listed in table 5.11.

Group	size (wds)	Group	size (wds)	Group	size (wds)
alt	6,492,546	eduni	6,040	rec	7,090,111
bionet	88,767	eunet	28,985	sci	1,252,309
bit	664,681	gnu	98,320	scot	7,208
biz	71,257	ieee	2,969	soc	2,796,821
chinese	635	info	2,106	talk	678,192
cogsci	9,589	k12	2,279	uk	539,271
comp	7,268,116	misc	1,195,578	vmsnet	16,752
ed	9,344	news	1,725,997		

Table 5.12: Sizes of the Usenet corpus group hierarchies.

Both full years of The Guardian were used, and with the King James version of the Bible, the Quran, the complete works of Shakespeare and Project Gutenberg constitute the style-guided portion of the total corpus. Project Gutenberg[1] is an online book initiative, and the corpus I have used consists of around 50 books and written material ranging from the American Constitution, through Mark Twain's "Adventures of Huckleberry Finn", to the CIA World Factbook.

The formal, but non style-guided, section of the whole corpus consists of the Leverhulme corpus, a transcribed set of secondary education students' essays, the Spoken English Corpus (as referred to in section 4.2) and the Philosophy corpus. The latter consists of several contemporary academic philosophical papers made available online by the IPPE project. None of these corpora are likely to have gone through a really formal editing phase, and hence they will not have encountered prescriptive style guides. However, since their authors have all been operating in a formal environment, they will all embody the idiosyncratic formal punctuation styles of their authors.

This contrasts with the final category of corpus material: free writing. This has been collected on-line from Usenet, the news mechanism of the InterNet.[2] The text in this corpus takes the form of (mainly) spontaneous (mainly) short messages from people all over the world, and since the environment is probably as informal as writing can get, the stricter conventions and notions of punctuation do not really apply. This illustrates the totally unconstrained way in which many people would write in the absence of style restrictions, and hence is likely to give a valuble insight into idiosyncratic usages, and the way that people are likely to write in informal media such as personal letters. Hence this corpus could be viewed as the most accurate and natural representation of punctuation use, since no formal style and form restrictions have been imposed on the writing.

The Usenet corpus has been divided into four subgroups for purposes of manageability. The *alt* (alternative newsgroups to the older 'official' hierarchies), *rec* (recreational topics) and *soc* (social topics) hierarchies of newsgroups have been analysed separately, since they are the

[1] Note that the Project Gutenberg corpus consists of all the works in that Project up to June 1994, barring repetitions and those files which are purely data, such as the US Census returns and π to 1,000,000 decimal places!

[2] I would like to acknowledge the assistance of Steve Finch, and his `grab` program, to extract text from Usenet spool files in collecting this corpus.

Figure 5.3: The punctuational unsuitability of programming languages

three largest, and the remaining hierarchies have been grouped together. The size difference of the various hierarchies can be seen in table 5.12. The reason why the largest hierarchy, *comp* (for computational topics), has not been included in the processing is that it included a great deal of non-textual material, such as graphics, computer code and (deceptively linguistic) programs, which would confuse the results, c.f. figure 5.3 (Adams, 1996).

Quantitative results

The numerical frequencies of occurrence of all the non-alphanumeric symbols in all the corpora used in this study are reported in tables 5.13 and 5.14. Note that these symbols are not necessarily isomorphic to punctuation marks. Some symbols may turn out not to be formal marks of punctuation, and some may have to combine to form these marks. For examples, a double quotation-mark can be composed of two apostrophe-like symbols, and the ellipsis is composed of three consecutive dots. The reference to dots in the results is for this reason, since not all the dots are necessarily full-stops. An additional comment on the punctuation symbols is that only those symbols that display have been reported, thus various control characters and sequences (such as the one to control the pound sign) are omitted. The frequencies of these characters were fairly low in the corpora though, so their omission does not affect the results unduly.

The tables in this section have been laid out to subdivide the punctuation symbols into three categories. The top category, for those symbols that can finish sentences, contains the dot and stress markers. The second category, for what we have been regarding as sentence-internal point punctuation, includes the other common point punctuation marks. The third category includes various methods of quotation and bracketing, and the fourth category contains all the novel and unusual punctuation symbols not usually regarded as forming part of the punctuation system.

Several observations can immediately be made from tables 5.13 and 5.14:

- There seems to be a profusion of punctuation symbols, over and above the number and variety of punctuation marks reported in the pilot study. While it is questionable that all of these marks are treatable as true punctuation — the mathematical symbols, for example — the majority of the symbols can be classed as legitimate punctuation c.f. the discussion on the nature of punctuation in the introduction

Punc. mark	Guardian 1990	Guardian 1991	Lever-hulme	Bible	Quran	Shake-speare	SEC
.	1,071,653	985,914	16,450	26,202	11,191	38,063	2,717
?	26,071	27,396	414	3,297	882	10,577	203
!	4,313	6,534	262	313	906	9,238	64
,	1,249,988	1,128,490	11,469	70,573	8,184	89,917	3,805
:	65,021	65,769	398	12,696	1,978	15,243	379
;	46,251	40,422	426	10,139	2,854	16,979	179
-	13,903	63,541	1,390	2	459	2,396	218
(	65,867	56,940	1,332	221	2,071	522	51
)	65,956	57,468	1,384	221	2,071	669	51
[	1,513	2,144	11			5,657	487
]	1,528	2,145	8			6,241	487
<	7		21				
>	4	4,869	33				
{	4		20		1		
}	4						
‘	323,143	108,859	1			6,910	
’	349,618	540,522	1,691	227	20		783
”	79	68	2,396		8		80
#	10	27	6				
*	519	977	39				
^	519				1		
|	1,430		2		1	618	
/	881	747	101		1		6
\			3				
_		216	80				
&	3,984	3,075	46			71	2
%	954	1,785	174				2
$	6,756	5,831	1				2
+	221	130	231				6
=	94	102	68		1		1
@		4	4				
~			7		9		
TAB			6,414				
Total	3,300,311	3,103,975	44,882	123,891	30,638	203,101	9,523

Table 5.13: Punctuation symbol frequencies in the corpora.

Punc. mark	Philo-sophy	Project Gutenberg	Usenet alt	Usenet rec	Usenet soc	Usenet rest
.	31,487	611,451	421,462	453,910	187,102	434,118
?	1,049	56,240	47,868	49,353	21,457	37,623
!	198	43,229	43,165	41,930	13,509	24,705
,	31,669	1,068,838	316,649	343,151	139,909	319,553
:	3,959	141,589	82,969	76,628	29,034	90,994
;	1,443	159,506	23,234	13,948	4,773	29,198
-	11,136	103,431	160,556	201,894	49,912	190,140
(	7,154	74,971	68,847	117,032	30,734	85,870
)	7,452	77,545	77,646	136,391	36,161	97,134
[	530	8,385	10,772	23,902	3,739	9,105
]	467	8,521	10,536	24,133	4,690	10,223
<	57	2,250	15,184	14,632	5,571	14,121
>	370	2,580	31,060	28,369	11,817	26,397
{	42	5,339	4,026	902	505	2,711
}	58	5,129	4,180	941	295	2,582
‘	661	16,219	10,051	5,049	6,960	6,326
’	1,697	39,395	25,830	18,954	15,287	21,341
”	10,560	225,984	102,587	84,868	41,538	78,544
#	749	16,545	16,016	15,518	4,417	16,620
*	1,248	22,117	38,954	53,296	18,181	88,451
^	58	142	16,451	20,598	22,302	12,103
|	177	1,093	11,383	15,488	6,495	10,198
/	237	2,308	15,075	16,957	6,218	15,892
\	150	90	8,241	8,601	2,923	5,356
_	3,091	42,337	45,868	40,538	14,896	36,076
&	457	354	6,307	6,363	1,074	6,955
%	65	83,540	6,702	5,690	2,321	9,966
$	29	10,874	12,126	24,390	3,165	12,464
+	46	7,952	16,014	20,601	8,177	23,476
=	872	26,975	32,886	31,796	12,534	35,367
@	7	362	12,136	3,847	3,120	3,099
~	32	329	7,422	4,262	6,965	3,821
TAB	592					
Total	120,799	2,865,620	1,702,203	1,903,932	715,781	1,757,829

Table 5.14: Punctuation symbol frequencies in the corpora.

- The spread of symbols seems to vary considerably between the corpora. The Usenet subcorpora, Project Gutenberg, the Philosophy papers and the Leverhulme corpus all seem to contain instances of almost all the punctuation symbols.[3] Whilst the two years of The Guardian seem to contain at least some instances of most of the symbols, the Bible, Quran, Shakespeare and SEC have very poor coverage of all but the most common symbols. Maybe this serves to indicate that these symbols are the key ones, the ones that are necessary for full understanding of the written text, and that the others are just devices to avoid making points in a clearer, more linguistic, fashion?

- The Usenet subcorpora and Project Gutenberg seem to use the less frequent punctuation symbols (those less commonly associated with punctuation marks, in the bottom section of the tables) a great deal more frequently than any other corpus. This could be due to the freer, lazier writing style mentioned above, or due to different material. It is unlikely that the e-texts in Project Gutenberg are very ‘free’ as regards punctuation, but since some rather less linguistic work is included (InterNet guides, and the CIA factbooks) it could be the case that this material contains rather more of the non-usual punctuation symbols. Those Usenet newsgroups which fall into the technical domain could also be demonstrating this difference in style, with greater use of the less linguistic punctuation symbols.

- Some stylistic differences emerge. It is clear from the figures that, for example, The Guardian uses single quotation marks (which as it happens are doubled up to form the double quotes) rather than the double-quote symbol, and that this situation is reversed in many of the Usenet texts.

- The most popular punctuation symbols remain the dot (we cannot yet determine whether all instances of the dot are full-stops) and the comma. There is no clear third place symbol — it varies between colon, semicolon, dash and closing quote between all the corpora.

- The opening and closing single quotation marks are rarely balanced. This is due not only to some of the corpora using quotation conventions that use the closing single-quote for both opening and closing purposes, but also due to the potential confusion of some word-final sub-lexical apostrophes with word-final closing quotation marks.

Comparative results

Interesting as the previous set of results are, they are only really of use for the study of the magnitude of the data themselves. It is difficult to compare the results between the different corpora, since they are of such different sizes. Therefore tables (5.15) and (5.16), which give the frequency of each symbol as a percentage of the total number of punctuation symbols in that corpus, are of more use.

The observations that can be made from tables 5.15 and 5.16 are as follows:

[3] The only reason the Usenet subcorpora are missing the TAB is that the extraction program replaces these with spaces.

Punc. mark	Guardian 1990	Guardian 1991	Lever-hulme	Bible	Quran	Shake-speare	SEC
.	32.5	31.8	36.7	21.1	36.5	18.7	28.5
?	.8	.9	.9	2.7	2.9	5.2	2.1
!	.1	.2	.6	.3	3.0	4.5	.7
,	37.9	36.4	25.6	57.0	26.7	44.3	40.0
:	2.0	2.1	.9	10.2	6.5	7.5	4.0
;	1.4	1.3	.9	8.2	9.3	8.4	1.9
-	.4	2.0	3.1	.0	1.5	1.2	2.3
(	2.0	1.8	3.0	.2	6.8	.3	.5
)	2.0	1.9	3.1	.2	6.8	.3	.5
[	.0	.1	.0			2.8	5.1
]	.0	.1	.0			3.1	5.1
<	.0		.0				
>	.0	.2	.1				
{	.0		.0		.0		
}	.0						
‘	9.8	3.5	.0			3.4	
’	10.6	17.4	3.8	.2	.1		8.2
”	.0	.0	5.3		.0		.8
#	.0	.0	.0				
*	.0	.0	.1				
^	.0				.0		
|	.0		.0		.0	.3	
/	.0	.0	.2		.0		.1
\			.0				
_		.0	.2				
&	.1	.1	.1			.0	.0
%	.0	.1	.4				.0
$	.2	.2	.0				.0
+	.0	.0	.5				.1
=	.0	.0	.2		.0		.0
@		.0	.0				
~			.0		.0		
TAB			14.3				
Total	99.8	100.1	100.0	100.1	100.1	100.0	99.9

Table 5.15: Percentage of total punctuation accounted for by each symbol.

Punc. mark	Philo-sophy	Project Gutenberg	Usenet alt	Usenet rec	Usenet soc	Usenet rest
.	26.1	21.3	24.8	23.8	26.1	24.7
?	.9	2.0	2.8	2.6	3.0	2.1
!	.2	1.5	2.5	2.2	1.9	1.4
,	26.2	37.3	18.6	18.0	19.5	18.2
:	3.3	4.9	4.9	4.0	4.1	5.2
;	1.2	5.6	1.4	.7	.7	1.7
-	9.2	3.6	9.4	10.6	7.0	10.8
(	5.9	2.6	4.0	6.1	4.3	4.9
)	6.2	2.7	4.6	7.2	5.1	5.5
[	.4	.3	.6	1.3	.5	.5
]	.4	.3	.6	1.3	.7	.6
<	.0	.1	.9	.8	.8	.8
>	.3	.1	1.8	1.5	1.7	1.5
{	.0	.2	.2	.0	.1	.2
}	.0	.2	.2	.0	.0	.1
‘	.5	.6	.6	.3	1.0	.4
’	1.4	1.4	1.5	1.0	2.1	1.2
”	8.7	7.9	6.0	4.5	5.8	4.5
#	.6	.6	.9	.8	.6	.9
*	1.0	.8	2.2	2.8	2.5	5.0
^	.0	.0	1.0	1.1	3.1	.7
|	.1	.0	.7	.8	.9	.6
/	.2	.1	.9	.9	.9	.9
\	.1	.0	.5	.5	.4	.3
_	2.6	1.5	2.7	2.1	2.1	2.1
&	.4	.0	.4	.3	.2	.4
%	.1	2.9	.4	.3	.3	.6
$	.0	.4	.7	1.3	.4	.7
+	.0	.3	.9	1.1	1.1	1.3
=	.7	.9	1.9	1.7	1.8	2.0
@	.0	.0	.7	.2	.4	.2
~	.0	.0	.4	.2	1.0	.2
TAB	.5					
Total	100.2	100.1	99.7	99.6	100.1	100.2

Table 5.16: Percentage of total punctuation accounted for by each symbol.

- While the comma and dot are the two most popular marks their placings and relative frequency vary greatly. In the corpora that were designated as style-guided, the comma occurs more frequently than the dot, occurring from 4.6% more frequently (The Guardian, 1991) to 35.9% more frequently (the King James Bible, where there are roughly three times more commas than dots), where these percentages refer to the total of all punctuation symbols observed. The only exception to this is the Quran, where there are almost 10% fewer commas, but this is because the sentences in the Quran have a far simpler structure than those of any other corpus, c.f. the lowest average sentence length in table 5.11.

  The formal, but non style-guided, corpora produced more varied results: The SEC has 11.5% more commas than dots, the Philosophy corpus is almost at parity, with just 0.1% difference, and the Leverhulme corpus has 11.1% fewer commas than dots. The informal, freely produced Usenet corpus has more dots than commas in all of its subcorpora, the difference lying at around 5% or 6% in all cases.

  This suggests that formal writing produces more convoluted sentences than free writing does, either for stylistic reasons, because the concepts that are being communicated are more complex, or maybe just because the more informal writing styles miss out punctuation marks that the formal styles would include. The more complex a sentence is, the more internal punctuation marks it is likely to contain, and since the comma is the most common sentence-internal punctuation mark, it becomes the most popular symbol in complex text.

- Stress markers are still relatively infrequent, compared with normal dots. They are least frequent in the official style-guided material (e.g. The Guardian) and are most frequent in the informal material, such as Usenet. Interestingly enough, however, the Bible, the Quran and Shakespeare all have very high percentages of stress markers, particularly Shakespeare. Presumably this is due to high emotive content of these works!

- Of the most infrequent set of punctuation symbols, those which occur most frequently and that are not mathematical or currency-related are the underscore and asterisk symbols, in the Usenet corpora, Project Gutenberg and the Philosophy papers.

- The matching symbols seem, for the most part, to be paired in terms of frequency. Minor discrepancies can be explained by missed openings or closings (either by the author or by the analysis software), but the larger discrepancies will be due to one of the symbols being used for another purpose in that corpus other than its matching role (as discussed previously, with various non-standard quotation conventions, and confusion with the apostrophe).

- The symbols that correspond to the remaining items of point punctuation (colon, semi-colon and dash) also occur with greater frequency in most of the corpora than the unusual symbols. In some of the corpora, however, one or other of these symbols will have been neglected, for example the dash in the King James Bible, where due to presumably stylistic considerations there are only two dashes in the entire work. This practice seems to have been specific to the Bible (c.f. the quotation at the start of chapter 6) rather than

as a result of historical practice. For information, the two dash symbols are combined together to form an en-dash, so that there is actually only one dash in the whole King James Bible, in Exodus, chapter 32, verse 32 (5.17). Even then, as can be seen, the dash occurs in combination with another punctuation mark.

(5.17) [Exodus 32:32] Yet now, if thou wilt forgive their sin–; and if not, blot me, I pray thee, out of thy book which thou hast written.

Punctuation distribution

The data, as presented so far, yields interesting insights into the relative amount of punctuation in a corpus, but still does not really reveal anything about the distribution of these punctuation symbols. Calculating the average quantities of each symbol that can be expected to occur in every sentence is a more useful measure, and these quantities are tabulated in tables 5.17 and 5.18.

Some interesting observations from the examination of these tables follow:

- Some corpora appear to be able to have more than one dot per sentence. There are several explanations for this: these dots could be compounded into ellipses (…) or the sentence could in fact be containing several sub-sentences, with dots appearing in brackets or quotation marks, for example.

- Some corpora appear to have sentence-stopping characters (dot and stress-markers) whose cumulative frequency is less than one per sentence. This only happens with the Quran and Shakespeare, so it could be the case that the last sentences of verses or paragraphs are unstopped.

- Where the cumulative frequency of sentence-stopping characters is greater than one per sentence, there are two possibilities. Either the marks are being conjoined (!!!) or they are occurring within sentences. For dots this is as mentioned previously, but stress markers can occur legitimately within normal sentences, without having a stopping role.

- The comma seems, on average, to be the most popular symbol. In most corpora it is likely to occur at least once per sentence (up to the maximum of almost 2.5 times on average per sentence in the Bible), except for in the Leverhulme corpus, where it only occurs with a likelihood of .75, the Usenet corpora, where the likelihood is around .67, and the Quran, where only just over half the sentences are likely to contain a comma. This seems, again, to reflect the style of writing employed in the formal and informal corpora.

- The only non point symbols that seem to occur with any regularity are the matched brackets and various quotation symbols (the precise identity of which varies between corpora), the asterisk and underscore in the corpora mentioned previously, and one-offs such as the percent sign in Project Gutenberg (an artifact of the inclusion of the CIA world factbooks?), the caret in one of the Usenet subcorpora (soc) and the TAB character in the Leverhulme corpus.

Punc. mark	Guardian 1990	Guardian 1991	Lever-hulme	Bible	Quran	Shake-speare	SEC
.	1.11	1.12	1.06	.87	.78	.60	.93
?	.03	.03	.03	.11	.06	.17	.07
!	.00	.01	.02	.01	.06	.15	.02
,	1.30	1.28	.74	2.35	.57	1.41	1.30
:	.07	.07	.03	.42	.14	.24	.13
;	.05	.05	.03	.34	.20	.27	.06
-	.01	.07	.09	.00	.03	.04	.07
(	.07	.06	.09	.01	.14	.01	.02
)	.07	.07	.09	.01	.14	.01	.02
[	.00	.00	.00			.09	.17
]	.00	.00	.00			.10	.17
<	.00		.00				
>	.00	.01	.00				
{	.00		.00		.00		
}	.00						
‘	.34	.12	.00			.11	
’	.36	.61	.11	.01	.00		.27
”	.00	.00	.15		.00		.03
#	.00	.00	.00				
*	.00	.00	.00				
^	.00				.00		
|	.00		.00		.00	.01	
/	.00	.00	.01		.00		.00
\			.00				
_		.00	.01				
&	.00	.00	.00			.00	.00
%	.00	.00	.01				.00
$	.01	.01	.00				.00
+	.00	.00	.01				.00
=	.00	.00	.00		.00		.00
@		.00	.00				
~			.00		.00		
TAB			.41				
Total	3.43	3.53	2.89	4.13	2.14	3.19	3.25

Table 5.17: Average number of punctuation symbols per sentence.

Punc. mark	Philo-sophy	Project Gutenberg	Usenet alt	Usenet rec	Usenet soc	Usenet rest
.	1.09	.94	.91	.87	.91	.92
?	.04	.09	.10	.09	.10	.08
!	.01	.07	.09	.08	.07	.05
,	1.09	1.65	.69	.66	.68	.68
:	.14	.22	.18	.15	.14	.19
;	.05	.25	.05	.03	.02	.06
-	.38	.16	.35	.39	.24	.40
(	.25	.12	.15	.23	.15	.18
)	.26	.12	.17	.26	.18	.21
[	.02	.01	.02	.05	.02	.02
]	.02	.01	.02	.05	.02	.02
<	.00	.00	.03	.03	.03	.03
>	.01	.00	.07	.05	.06	.06
{	.00	.01	.01	.00	.00	.01
}	.00	.01	.01	.00	.00	.01
‘	.02	.02	.02	.01	.03	.01
’	.06	.06	.06	.04	.07	.05
”	.36	.35	.22	.16	.20	.17
#	.03	.03	.03	.03	.02	.04
*	.04	.03	.08	.10	.09	.19
^	.00	.00	.04	.04	.11	.03
|	.01	.00	.02	.03	.03	.02
/	.01	.00	.03	.03	.03	.03
\	.01	.00	.02	.02	.01	.01
_	.11	.07	.10	.08	.07	.08
&	.02	.00	.01	.01	.01	.01
%	.00	.13	.01	.01	.01	.02
$	.00	.02	.03	.05	.02	.03
+	.00	.01	.03	.04	.04	.05
=	.03	.04	.07	.06	.06	.08
@	.00	.00	.03	.01	.02	.01
~	.00	.00	.02	.01	.03	.01
TAB	.02					
Total	4.17	4.41	3.68	3.66	3.48	3.73

Table 5.18: Average number of punctuation symbols per sentence.

- The average number of punctuation symbols per sentence varies between corpora, from a minimum of 2.14 in the Quran (a corpus that we have already shown to be stylistically simple, it still has over two symbols per sentence!) to a maximum of 4.41 in Project Gutenberg. Since this effectively means that every sentence of the English language is likely to have between 2 and 5 symbols of punctuation in it, the argument for studying punctuation and including it in computational linguistics systems is greatly enhanced.

Punctuation regularity

The problem with the results above is that they are not really comparable across the corpora. Knowing that a certain symbol is likely to appear in half of the sentences of one corpus and in every sentence of another is not useful unless some information is known about the relevant sentence lengths. It could be the case, for example, that the sentences in the second corpus are twice as long, and hence usage of the punctuation symbol is the same. Therefore the above data were retabulated to show the average number of words that occur between each instance of every punctuation symbol in the corpora, and the new results are shown in tables 5.19 and 5.20.

From these tables it now emerges that:

- The small number of commas per sentence in the Quran was an anomaly of the short sentence length. In fact there are comparable distances between commas in The Guardian and the Quran, and the only real anomalies are Shakespeare, where there are only ten words on average between commas, Project Gutenberg, at thirteen words, and the Leverhulme corpus, where the distance is a long thirty-one words.
- The stress markers ? and ! are used far more frequently in the Usenet corpora than in the others. Typically, overuse of these symbols has been ascribed to 'poor' style and bad writing, but does the high frequency in Usenet mean that such use of these symbols is the real way that the modern language operates, and that the formal, style-guided environments are artificial?
- The dash symbol also seems to be used to a far greater extent in Usenet than in the other corpora, and the same is true of the colon, where the frequency of use in Usenet is only matched by that in Bible, Quran, Shakespeare and Project Gutenberg. The semicolon is used less than the colon in the four subcorpora of Usenet, but still more frequently in certain Usenet subcorpora than in The Guardian. It has long been argued that the dash is used far more in free, 'colloquial' writing than in formal settings, but the results for semicolon and colon are surprising, since the prevalent attitude has always been that these symbols are mis- and under-used. This study shows that not to be the case (except of course in the Leverhulme corpus, but since that is produced entirely by secondary school children, whose written language could be argued not yet to be perfected, it is a result that can be ignored or at least bypassed).
- The average number of words between (any) punctuation symbols is not constant across the corpora. It seems to decrease from the formal corpora to the less formal ones. There

Punc. mark	Guardian 1990	Guardian 1991	Lever-hulme	Bible	Quran	Shake-speare	SEC
.	22	22	22	31	15	25	20
?	919	790	859	249	191	89	269
!	5,556	3,311	1,357	2,622	186	102	853
,	19	19	31	12	21	10	14
:	369	329	893	65	85	62	144
;	518	535	835	81	59	55	305
-	1,724	341	256	410,366	367	392	251
(	364	380	267	3,714	81	1,799	1,071
)	363	377	257	3,714	81	1,404	1,071
[	15,838	10,093	32,327			166	112
]	15,683	10,088	44,449			150	112
<	∞		16,933				
>	∞	4,444	10,776				
{	∞		17,780		168,663		
}	∞						
‘	74	199	355,594			136	
’	69	40	210	3,616	8,433		70
”	303,336	318,220	148		21,083		683
#	∞	801,443	59,266				
*	46,173	22,148	9,118				
^	46,173				168,663		
|	16,758		177,797		168,663	1,520	
/	27,200	28,968	3,521		168,663		9,102
\			118,531				
_		100,180	4,445				
&	6,015	7,037	7,730			13,228	27,307
%	25,119	12,123	2,044				27,307
$	3,547	3,711	355,594				27,307
+	108,432	166,454	1,539				9,102
=	254,931	212,147	5,229		168,663		54,614
@		∞	88,899				
~			50,799		18,740		
TAB			55				
Total	7	7	8	7	6	5	6

Table 5.19: Average number of words occurring between each instance of each symbol. (∞ represents numbers above 1,000,000)

Punc. mark	Philo-sophy	Project Gutenberg	Usenet alt	rec	soc	rest
.	16	22	15	16	15	15
?	494	244	136	144	130	170
!	2,617	318	150	169	207	259
,	16	13	21	21	20	20
:	131	97	78	93	96	70
;	359	86	279	508	586	219
-	47	133	40	35	56	34
(	72	183	94	61	91	75
)	70	177	84	52	77	66
[	978	1,640	603	297	748	703
]	1,110	1,613	616	294	596	626
<	9,090	61,104	428	485	502	453
>	1,400	5,328	209	250	237	242
{	12,337	2,575	1,613	7,860	5,538	2,361
}	8,933	2,680	1,553	7,535	9,481	2,479
‘	784	848	646	1,404	402	1,012
’	305	349	251	374	183	300
”	491	61	63	84	67	81
#	6,923	831	405	457	633	385
*	415	622	167	133	154	72
^	8,933	96,812	395	344	1,257	529
|	2,927	12,578	570	458	431	628
/	2,186	5,956	431	418	450	403
\	3,454	152,749	788	824	957	1,195
_	168	325	142	175	188	177
&	1,134	38,834	1,029	1,114	2,604	920
%	7,971	165	969	1,246	1,205	642
$	17,867	1,264	535	2,917	884	514
+	11,264	1,729	405	344	342	273
=	594	510	197	223	223	181
@	74,020	37,976	535	1,843	8,967	2,065
~	16,192	41,785	875	1,664	402	1,675
TAB	875					
Total	4	5	4	4	4	4

Table 5.20: Average number of words occurring between each instance of each symbol.
(∞ represents numbers above 1,000,000)

is a generalised belief that is widely held that people writing freely use too much punctuation, which this data would seem to support. However, since we have established that the common view is that people use too few colons and semicolons, the only candidates for this increased punctuation are the comma and dash. The comma is used slightly less frequently in the Usenet corpora than in the formal corpora, so it is unlikely to be the culprit. We have acknowledged that the dash is more prevalent, but at an occurrence rate of one every 40 words is unlikely to affect the overall punctuation statistics much. Therefore, it must be the case that this extra punctuation is coming from all symbols, in which case punctuation is either overused across the board (which seems a little strained as a hypothesis) or this use is the emergent standard and formal style-guided material is still using too little punctuation. The alternative is that formal material has been planned more, and therefore may have better structure, so that it does not need as much punctuation as spontaneously produced writing.

- One thing that tables 5.19 and 5.20 do illustrate very well is that certain of the punctuation symbols really do occur very infrequently indeed. If a particular symbol only occurs once every 10,000 words, then a natural language processing system that does not include a treatment of that symbol is unlikely to be unduly affected if it encounters that symbol, regardless of how high the syntactic or semantic content of that symbol is. Therefore there is an argument for a two-tier treatment of punctuation. These tiers would consist of a core group of symbols which should be implemented on every system, and that group of symbols which occur sufficiently infrequently for their omission not to have too adverse an effect on the performance of a system. This division will obviously change depending on the corpus that the data for the processing system is taken from, indeed the whole set of symbols may change from those represented in the tables here. It may be possible to find a corpus which uses other symbols, possibly from another character set, to signal punctuational functions. An example of this could be control-codes that determine things such as underlining or italicisation in word processors (c.f. the discussion of super-lexical punctuation in the introduction).

  The core group of punctuation symbols, based on the data collected here, should always consist of the seven point-type symbols at the top of the tables shown and whatever method(s) of bracketing and quotation that are employed in a particular corpus or tet-source. Specific other characters could then be added according to the demands of particular corpora.

Punctuation patterns

Now that the data have been examined for the quantity and frequency of punctuation symbols in texts, it is instructive to look at the patterns that punctuation occurs in. The 50 most common patterns in each corpus have been tabulated in frequency order in tables 5.21 and 5.22.
Several things can be observed from these tables:

- The mistakes — the twelfth line in the Quran column has an æ character in it. This was actually ctrl-Z in the original text, but has been reproduced incorrectly here. Similarly, the

Guardian-90	Guardian-91	Leverhulme	Bible	Quran	Shakespeare	SEC
e.	e.	e.	e,e.	e.	e.	e.
e,e.	e,e.	e,e.	e,e,e.	e,e.	⇒[e]	e,e.
e,e,e.	e,e,e.	⇒e.	e.	e!	e,e.	e,e,e.
e,e,e,e.	e,e,e,e.	e,e,e.	e,e,e,e.	e,	e?	e,e,e,e.
e,e,e,e,e.	e,e,e,e,e.	⇒e,e.	e:e.	e?	e!	[e]
e?	e?	e,e,e,e.	e?	e,e,e.	e,e!	e,e,e,e,e.
“e,”e.	e:e.	⇒e,e,e.	e,e,e,e,e.	e:e!	e,e?	[e:e]
e:e.	”e,”e.	e-e.	e,e?	e;e.	e,e,e.	e?
e,e,e,e,e,e.	e,e,e,e,e,e.	e	e,e:e.	e:e.	e⇒e.	’e,’e.
e“e”e.	e-e.	e"e"e.	e;e.	e;e,e.	⇒e	e,e,e,e,e,e.
ee.	.	⇒e,e,e,e.	e,e:e,e.	e:e?	e	e!
“e.”	e”e”e.	e(e)e.	e,e;e.	æe,e,e.	⇒e?	[e]e,e.
e“e”.	e;e.	e,e,e,e,e.	e:e,e.	e:e,e.	⇒e.	[e]e?
e;e.	e-e-e.	e’e’e.	e,e,e?	e,e;e.	e,e⇒e.	e;e.
.	”e.”	e;e.	e,e,e:e.	e(e)e.	e⇒e?	e:e.
e’e.	“e,”e.	(e)	e,e,e,e,e,e.	e,e?	e,⇒e.	e,e?
e(e)e.	e’e.	e⇒e.	e;e,e.	e	⇒e!	[e]e.
e,e“e”e.	e(e)e.	e?	e,e,e;e.	e,e,e,e.	e;e.	[e][e][e:e][e:e,e,e]
“e,e,”e.	e”e”.	⇒e-e.	e,e,e:e,e.	e;	⇒e,e.	’e?’
e,e,e,e,e,e,e.	e:e,e.	e:e.	e,e;e,e.	e,e,	⇒e,e!	e;e,e.
e:e,e.	e,e?	e,e-e.	e,e,e,e?	e,e;e,e.	⇒[e,e]	e:e,e.
e,e?	e”e”e.	e(e).	e,e,e,e,e,e,e.	e(e)e,e.	e,e;e.	e...
e,e,e“e”e.	e,e,e,e,e,e,e.	e(?)	e,e:e,e,e.	e:	e,e,e!	’e.’
e“e”e,e.	-e.	e%e.	e:e,e,e.	e,e:e.	e,e,⇒e.	e-e.
e,ee.	e,e-e.	e-e,e.	e,e,e,e:e.	e,e!	e,e,e?	e,e;e.
“e,e.’	’ ..e.-	e,e,e,e,e,e.	e;e,e,e.	e,e;	⇒e,e?	e:
“e,”e,e.	ee.	e"e".	e,e,e;e,e.	e(e).	e⇒e,e.	’e,e.’
e,e,e“e”.	“e.”	e-e-e.	e,e,e,e;e.	e,e(e).	e,e,e⇒e.	e:e,e,e.
“e.e.”	e:e,e,e.	e’e.	e,e,e,e:e,e.	e,e:	e⇒e!	e,e:e.
e:“e.”	e	⇒e⇒e.	e,e,e,e,e?	e,e,e;e.	⇒e,⇒e.	e,e...
e,e“e”.	”e,e,”e.	e’e’.	e,e,e:e,e,e.	e:e;e.	e:e.	e,e,e,e,e,e,e.
e	e,e,e”e”.	⇒⇒e.	e,e;e,e,e.	e:e,e,e.	⇒[e,e,e]	[e]e,e,e.
e,e,ee.	e“e”.	⇒e(e)e.	e,e,e,e,e:e.	e,e(e)e.	e;e⇒e.	[e]e,e?
ee,e.	”e,”e,e.	e+e.	e,e:e;e.	e,e,e,e,e.	e;⇒e.	e,e:e,e.
e“e.”	e,e”e”e.	⇒	e;e:e.	e(e)e,	e,e;e⇒e.	e,e;e,e,e.
e(e).	e,e”e”.	e⇒⇒e.	e,e:e,e,e,e.	e,e:e!	⇒e⇒e.	e,e,e;e.
e:e,e,e.	e,e:e.	⇒e,e,e,e,e.	e:e,e,e,e.	e,e	e,e,e,e.	e;e,e,e,e.
e:e(e).	e,e,e?	e⇒e⇒e.	e,e,e,e,e,e,e,e.	e:e,e?	(e:)	e:e,e,e,e.
eee.	e(e).	e:	e;e,e,e,e.	e(e)e,e,e.	e,e;⇒e.	e,e,e;e,e.
e,e,e?	”e,e.”	e,e;e.	e,e,e,e,e;e.	e;e;e.	⇒e,e,e.	e,e,e,e,e?
e;e,e.	e,e’e.	e.⇒	e,e!	e(e)e;e.	e,e⇒e?	e’e’e.
e,e’e.	e,e,e”e”e.	e!	e,e,e,e;e,e.	e,e:e?	e,e:e.	e’e’.
e,e:e.	e:e(e).	⇒e’e’e.	e,e;e:e.	e;e,e,e.	e::⇒e.	e,e-e.
e,e(e)e.	e,e(e)e.	e,e?	e,e,e,e,e:e,e.	e:e;e,e.	e:e⇒e.	e,e,e:e.
e’e,e.	”e.e.”	e(e)e,e.	e,e,e,e,e,e,e,e,e.	e,e,e?	e⇒e,e?	e,e’e’.
e,e;e.	e;e,e.	⇒e?	e;e:e,e.	e,e;e;e.	e⇒e,e⇒e.	e,e!
e-e.	e:”e.”	e,e(e)e.	e:e;e.	e,e;e,e,e.	e,–	’e,e,’e.
e,e	e,e,e-e.	e,e’e’e.	e,e,e,e,e,e?	e,e:e,e.	e,⇒e?	’e,’e,’e.’
e“e”,e.	e”e”e,e.	e,e"e"e.	e!	(e)e,e.	⇒e,e⇒e.	e;e,e,e.
e,e,e,e,e,e,e.	e’e,e.	e⇒⇒e⇒⇒e.	e;e;e.	e;e?	e:)	e-e,e.

Table 5.21: Top 50 punctuation patterns (⇒ replaces TAB)

Philosophy	Gutenberg	Usenet alt	Usenet rec	Usenet soc	Usenet rest
e.	e.	e.	e.	e.	e.
e,e.	e,e.	e,e.	e,e.	e,e.	e,e.
e	e,e,e.	e,e,e.	e	e,e,e.	e
e,e,e.	e,e,e,e.	e?	e,e,e.	e	e,e,e.
e,e	e	e	e?	e?	e?
e:e.	e?	e!	e!	e!	e:
e,e,e,e.	e,e,e,e,e.	e,e,e,e.	e(e)e:	e,e,e,e.	e,e,e,e.
e?	e;e.	e(e)e:	e...	e(e)e:	e:e.
(e)e.	"e,"e.	e...	e:	e:	e!
e,e,e,e,e.	"e?"	e:	e,e,e,e.	e...	e(e)e:
e,e,e	"e."	e,e?	e,e	.	e(e)e.
e(e)e.	e!	e:e.	e,e?	e,e?	e,e
e,e.,	e,e,e,e,e,e,e.	e"e"e.	e(e)e.	e<e>e:	e,e?
e"e"e.	e:e	e<e>e:	e:e.	e"e"e.	e...
(e).	e,e?	e(e)e.	e"e"e.	e(e)e.	e"e"e.
e:e,e.	"e!"	.	e(e).	e,	e:e
e(e).	e,e;e.	e,e!	.	e:e.	.
e:	e:e.	e,e,e,e,e.	e<e>e:	e,e	e(e).
e.,	e;e,e.	e,e	e,	e,e,e,e,e.	e-e.
e;e.	.	e,e...	e,e!	e-e.	e,e,e,e,e.
.	e,e;e,e.	e,	e,e...	e,e!	e<e>e:
e,e,e,e,e,e.	e,e!	e-e.	e-e.	e..	e,
(e)e:e.	"e,e."	e..	e,e,e,e,e.	e(e).	e:e,e.
e-e.	e,e,e,e,e,e,e,e.	e(e).	-e	e,e...	e:e?
&e,e.	-#_e:e	e,e,e?	e$e.	e,e,e?	e;e.
e,e"e"e.	"e,e,"e.	-e	e:e	e"e".	e,e:
e,e?	e:	e:e,e.	e..	-e	e,e!
e:e	e,e,e?	"e."	e:e,e.	e;e.	e,e(e)e.
e,	e,e,e;e.	e;e.	e,e(e).	e:e	e,e(e).
e"e."	e;e,e,e.	e....	e,e(e)e.	e,e"e"e.	e$e.
e,e(e).	e,e:e.	e:e	e;e.	e’e.	e,e,e
.e"e"e:e.	"e,e?"	e*e*e.	e,e,e	e,e(e)e.	-e
e-	*e,e	e"e".	e)e.	e:e,e.	e(e)
(e)e,e.	"e,"e,e.	e,e"e"e.	e,e"e"e.	[e]	e"e".
e"e"e,e.	e	e,e:	e,e,e?	e,e(e)e:	-e.
(e.)	:e	:)	-e.	e,e:	*e.
e,e(e)e.	e:e,e.	e,e,e,e,e,e.	e....	e,e,e,e,e,e.	e,e"e"e.
e(e),e.	e,e:	e>e.	e,e:	e*e*e.	e,e.,
e,e:e.	e:e(e)	e,e(e)e.	e!!	e!!	e,e.;
e"e"e	e,e:e,e.	e,e(e).	e,e(e)	e"e"e,e.	e)e.
e"e".	e,e;e,e,e.	e:e?	e"e".	e<e>e(e)e:	e)e?
e(e)e,e.	e,e,e;e,e.	-e.	:)	:-)	e,e...
e"e	"e,e!"	[e]	e(e)	(e)	e,e,e?
.e.e.e.e,e.	"e.e."	"e?"	e:e?	-e.	e.,,
.e.e,e.	e-e.	e,e(e)e:	e!!!	e....	e(e)e,e.
e!	-#_e:e(e)	e"e"e,e.	e)e:	e,e(e).	e,e,e,e,e,e.
e:e,e,e.	e,e,e,e,e,e,e,e.	e"e."	[e]	e.,	e:e:e
e,e-e.	e:e(e.)	e,e,e...	e(e),e.	e"e."	⊢⟶*e|e|⊢⟶*e(e)
e,e,e?	e,e,e:e.	(e)	e*e*e.	e)e:	e*e*e.
.e,e.	e;e,e,e,e.	e,e,e!	e(e)e,e.	e,e-e.	e:e,e,e.

Table 5.22: Top 50 punctuation patterns

lines in the Guardian columns that contain adjacent 'e's are actually separated by ASCII character 234, which is a pound symbol, but it has been reproduced incorrectly. This is an example of the problem of character sets and particular pieces of software. The symbol in the data was processed by the analysis software, and is displayable in certain cases, but not directly on the terminal or in the current font being used in this thesis.

- The standard order of complexity can be observed in the results. The more punctuationally complex the sentence (i.e. more clauses — more commas, etc.) the less frequently it occurs. Interestingly this pattern is broken in the Bible, where a plain, one-clause sentence is only the third most likely possibility, reflecting the complexity of the text.
- Evidence can be seen for the relative importance of the stress markers. Sentences with these marks appear high on the list for the Usenet corpora and Shakespeare and the Quran, but lower down or not at all for The Guardian and the Philosophy corpus.
- We can see those (usually compound) punctuation marks which cannot be spotted in the raw data alone. The ellipsis (…) appears prominently in the SEC and the Usenet corpus, and the so-called 'poor style' emphasising punctuation stress markers (!!, !!!) can be seen in the Usenet subcorpora. Whilst these may be poor style from a formal point of view, Parkes (1992) reminds us that they still convey information different to that of the single stress markers, and so they should be catered for, although maybe not in the 'core group' of punctuation marks discussed previously (c.f. section 10.6).
- The different styles of quotation emerge. The Guardian uses double quotes composed of inverted commas, whereas the Leverhulme corpus and Usenet use the keyboard double quote character. The SEC uses the unmatched inverted commas that signify the use of the keyboard single-quote character. It is also noticeable that the Quran, the Bible and Shakespeare seem not to contain any quotation at all in their top 50 entries.
- Unstopped sentences — several sentences are visible in the tables that illustrate the problem of unstoppedness. Some, as in those occurring in the first five rows of table (5.22) are simply left with dangling lexical items, where there is either a line missing, or the final punctuation has been omitted. Others, as in the patterns from the Quran, show that the sentence in question has been ended with an inappropriate punctuation mark: a comma, colon, or semicolon. This is not necessarily incorrect — consider the launching sentence of this itemised list, for example — but merely reflects a more structural use of written language. Until we can accurately represent and process white space and structure on a computer, such problems may never be resolved.
- We see that in The Guardian, for example, the matching of punctuation marks is not always performed correctly. The double quotes in some lines of the 1991 column, for example, are both of the closing variety. Somehow, the wrong type of quote has been inserted. Similarly in the 1990 column, there is an entry for the pattern *["e,e.']*, where the closing quote is of the wrong kind.

- We can see some new punctuation marks emerging, particularly in the Usenet corpora. Three of the four Usenet subcorpora include 'smilie' punctuation marks *[:) :-)]* which are used informally to indicate irony or gentle humour which might otherwise be lost, given the orthographic representation, in their 50 most frequent patterns, which suggests that certainly in the electronic text domain this mark should perhaps be included in a treatment of punctuation.
- Other specialised punctuation symbols also emerge. In Project Gutenberg, for example, *[_#_]* is used as a marker of some kind, and the matched asterisks in the Usenet corpora are most likely to be some sort of emphasisers for a particular word or phrase — the simplistic way of underlining something, or placing it in bold, where the only medium available is the ASCII character set.
- We can also see that punctuation symbols that appear to be punctuation marks are not necessarily so. Take for example the fourth from bottom entry from Shakespeare *[e,–]*. The dash in this entry is in fact composed of two dash symbols (and so represents an en-dash, by the composition concept which will be familiar to most LaTeX writers). If all the dashes in the corpus were composed like this, then the straight frequency report for the dash symbol would be misleading. Similarly, there exists a punctuation mark *[:-]* which is compounded of more than one symbol, which could also lead to inaccuracies in an analysis report. However, by manually examining the corpora, it is clear that usually dash symbols are only used individually, and that unusual combinations such as *[:-]* occur with very low frquency, so the numerical statistics reported previously will not have been unduly affected by the aggregation of punctuation symbols.
- Both types of Nunberg's (1990) quote transposition can be observed in the data. Sentences with final punctuation characters inside and outside quotation marks can be observed.

5.3 Summary

We have seen what a huge variety of punctuation symbols there are, and how they can combine together to form punctuation marks. Frequency analysis suggests, however, that there is a smaller core of symbols that accounts for the majority of punctuation, so it is with this set that initial investigations should be made. The important point is to realise that this set, and even the larger set reported in these results, is not exhaustive, for even if we run out of punctuation symbols, the existing ones can always combine in a new way.

Systems should not be overspecific when it comes to bracketing and quoting, both in their choice and treatment of the different characters, and their enforcing of the matching requirements. Indeed, in the case of quotation there exist phenomena, such as Victorian quotation, in which the closing mark is not necessarily present. (Victorian quotation refers to the practice of quoting extended portions of text, and putting an opening quotation mark at the start of each new quoted paragraph, but only to put a closing mark at the end of the whole quotation.) The

presence of word external apostrophes can also confuse the issue slightly as regards matching quotation.

Systems should also not always expect punctuation, even in places where it is usually mandatory. We have seen that there are a good many sentences in various texts that do not appear to have any sentence-final punctuation at all.

Nunberg's (1990) prescriptive points have been shown to be incorrect, in that the phenomena they refer to occur in real language. He suggests a ban on more than one nested colon-expansion, a situation that can occur in certain circumstances, particularly when the outer colon-expansion has a very complex structure. Also prescribing various complex symbols — the multiple exclamation marks, colon-dash — is not productive, as they occur frequently, and have their own meaning, in real language.

Some of Nunberg's nesting criteria are also inappropriate. Brackets can nest in natural language (although they tend not to) and quotes that occur within quotes do not necessarily need quotation marks that are differentiable from those of the surrounding quotation (although if such differentiation is available, it makes disambiguation easier — c.f. the quote from Nunberg (1990) in section 3.2).

This seems, to a certain degree, to be advocating the 'suck it and see' school of punctuation processing. Preconceived ideas of punctuation extent and variety should be avoided (although of course we want to codify some notions of likely punctuation form and function). The core group of punctuation symbols appear to be fairly invariant, and could safely be codified in any system — the capacity to handle, process and preferably ascribe sensible meaning to novel punctuation should also be present though, to give any natural language processing system a comprehensive treatment as regards the punctuation phenomenon.

To formalise, then, there seem to be two distinct sets of punctuation phenomena. The first set, consisting of various standard point punctuation marks and some system- or corpus-specific methods of quotation and bracketing (5.18), is the set that will contain the great majority of punctuation encountered. Furthermore, the marks in this set can be assigned a standard interpretation that will be invariant from system to system and corpus to corpus. The second set, however, will consist of those less frequent, novel punctuation marks whose occurrence and interpretation are system specific, e.g. (5.19). The variety of this set will therefore depend on the corpus to be processed, as will the specific interpretation that the marks contained in it should receive.

(5.18) **. , ; : — ( ) “ ” ? !**

(5.19) **@ # $ % ;-) &**

Numerically, it has emerged that every sentence of English text that is encountered by a processing system is likely to include between two and five punctuation marks, and that a punctuation mark of some variety is likely to be encountered from on average every fourth word to every seventh. The result that punctuation is so frequent in written English should emphasize the importance of developing an account whereby the information contained in those punctuation marks can be extracted and utilised in an analysis.

	journalism *formal*	novels *medium*	free writing *informal*	students' *learners*
freq. of ! and ?	LOW	high	HIGH	low
freq. of 2nd set	LOW		HIGH	
ratio of : to ;		=	HIGH	=
freq. of :		high	HIGH	LOW
ratio of . to ,		high	high	high
freq. of brackets			high	
freq. of dash			high	

Table 5.23: Genre-specific distinguishing punctuation.

Furthermore, it seems that punctuation usage might have some role to play in the process of genre-identification. Table 5.23 shows the punctuational features that seem to distinguish various genres of written material, as least so far as this analysis is concerned, so it would be interesting to see how far these results will generalise if these criteria were employed in a genre-identification system.

six The Syntactic Function of Punctuation

"The Bible bars the dash, which is the great refuge of those who are too lazy to punctuate. I never use it when I can possibly substitute the colon, and I save up the colon jealously for certain effects which no other stop produces. As you have no rules, and sometimes throw colons about with an unhinged mind, here are some rough rules for you."
(G B Shaw, in a letter to T E Lawrence)

Now that we have established the variety of punctuation marks that can occur, and seen how they can be divided into two distinct sets for the purposes of implementation, it is necessary to find out how a treatment of the set of standard marks (6.1) can be implemented. There are two obvious sides to this problem, which both need to be approached, the syntactic and the semantic. This chapter will deal with the former, namely the behaviour and function of punctuation in syntax, grammars and parsing.

(6.1) **. , : ; — ( ) " " ? ! ...**

The problem has typically been approached from the top down, so that the need has been seen for incorporating an account of one or more punctuation marks into an existing system. Thus the treatments that emerge tend to be rather *ad hoc*, embodying all the authors' idiosyncratic perceptions and ideas about punctuation rather than any theoretically or linguistically motivated notions. These accounts also tend to be rather performance driven: if incorporation of a particular punctuation mark does nothing to improve performance or efficiency in the system, it will typically be ignored (a good example is with cases where the treatment of the comma is omitted, because of the complexity of its use and the associated possibility for extra ambiguity).

To produce a properly motivated account of the function and performance of punctuation within syntax, it is necessary to start from the bottom up, i.e. to examine the way that the punctuation operates and is placed in the text, before integrating it into a syntactic processing

system. There are, essentially, two possible directions of approach — observational and theoretical — to try to determine the use of punctuation within syntax and enable conclusions and generalisations regarding this function of punctuation to be drawn.

6.1 Observational Approach

One of the best data sources for observational analysis of grammatical punctuation usage is a large, parsed corpus. It ensures a wide range of real language is covered, and because of its size it should minimise the effect of any errors or idiosyncrasies on the part of editors, parsers and transcribers. Unfortunately, these corpora are almost all hand-produced, so that some errors and idiosyncrasies are inevitable — one important preliminary part of the analysis is therefore to identify possible instances of these, and if they are clear, to remove them from the results. The set of parsed corpora is sadly very small but the corpora available are still sufficient to yield useful results.

What the current study will try to achieve is the extraction from parsed corpora of patterns of punctuation usage (for example, that a colon can occur between a syntactic entity of sentential type, and one of type 'noun phrase'). These will not only provide the patterns of usage of punctuation marks within the corpus, but will also hopefully illustrate some regularities in the usage of individual punctuation marks, which will allow a small set of generalised rules for punctuation placement to be synthesised.

The corpus chosen was the Dow Jones section of the Penn Treebank (size: 1.95 million words). The bracketings were analysed so that each 'node' that has a punctuation mark as one of its immediate daughters is reported, with its other daughters abbreviated to their syntactic categories, as illustrated in (6.2).

(6.2) a. [NP [NP the following] :] $\Longrightarrow$ [NP = NP :]
b. [S [PP In Edinburgh] , [S ...]] $\Longrightarrow$ [S = PP , S]
c. [NP [NP Bob] , [NP ...] ,] $\Longrightarrow$ [NP = NP , NP ,]

In this fashion each sentence was broken down into a set of such category-patterns, resulting in a set of different category-patterns for each punctuation symbol. These sets were then processed by hand to extract the underlying rule patterns from the raw category-patterns since these included aberrant phenomena such as instances of serial repetition (6.3) and 'lexical breakthrough' in cases where phrases are not marked in the original corpus (6.4).

(6.3) [NP = NP , NP , NP , NP or NP]

(6.4) [NP = each project , or activity PP]

The underlying rule-patterns thus produced represent all the ways that the specific punctuation marks behave in the corpus, and are good indicators of how the punctuation marks might behave in the rest of language. It should be possible to try to generalise these underlying rule-patterns into a small number of rules or principles that encapsulate the manner and syntactic contexts in which punctuation can occur.

Experimental Results

There were 12,700 unique category-patterns extracted from the corpus for the five most common marks of point punctuation, ranging from 9,320 for the comma to 425 for the dash. These rules were then reduced, by removal of serial repetitions and lexical breakthroughs, to just 137 underlying rule-patterns for the colon, semicolon, dash, comma, full-stop.

All these patterns were then checked against the original corpus to recover the original sentences. The sentences for patterns with low incidence and those whose correctness was questionable were carefully examined to determine whether there was any justification for a particular rule-pattern, given the content of the sentence — the pattern (6.5), for examples, occurring as the underlined portion of (6.6), seems not to be valid.

(6.5) [ADVP = PP , NP]

(6.6) The U.S. government in recent years has accused Japanese companies of excessively slashing prices on semiconductors and supercomputers — products Fujitsu and NEC make.

Taking the subset of rules relating to the colon, for example, shows that there are 27 underlying rule patterns from the original analysis, as shown in table 6.1.

By examining all (or at least a representative subset for very frequent rules) of the sentences in the original corpus that yield these underlying rule-patterns, the majority of them can be eliminated for a variety of reasons, as discussed below. The only real underlying patterns remaining are those shown in table 6.2.

The rest of the rule-patterns were eliminated because they represented idiosyncratic bracketings and category assignments in the original corpus, and so were covered by other rules. Some incorrect category assignments were made at the earlier data analysis stages, which explains why several of the revised rules have non-phrasal-level left-most daughters. Here are some examples of the inappropriate rule patterns.

NP=NP:NP	NP=S:NP	VP=VP:VP	S=S:S
NP=NP:PP	NP=PP:NP	VP=VP:NP	S=S:NP
NP=NP:VP	PP=PP:PP	VP=VP:PP	S=S:
NP=NP:S	PP=PP:	VP=VP:S	S=NP:S
NP=NP:	PP=AS IN:	VP=VP:	S=NP:VP
NP=NP:ADJP	PP=TO:	S=VP:NP	S=PP:S
NP=VP:NP		S=VP:S	S=IJ:S

Table 6.1: Underlying rule-patterns pertaining to the colon.

NP=NP:NP	NP=NP:S	NP=NP:PP	NP=NP:ADJP
PP=PP:PP	PP=P:NP	VP=V:S	VP=V:NP
S=S:S	S=S:NP	S=PP:S	S=VPING:NP

Table 6.2: Remaining colon rule-patterns

- S=NP:S — this pattern is inappropriate because the mother category should really be NP. Instances of this pattern in the corpus (6.7) are no different to instances of the similar rule with a NP mother and the pattern is more suited to a nominal interpretation. The problem has arisen in this case through confusion of sentential and top categories in the grammar. Almost all items in the corpus are marked as sentences, although not all fulfil that grammatical role.

  (6.7) Another concern: the funds' share prices tend to swing more than the broader market.

- NP=NP:VP — all the verb phrases that were daughters of the nodes for this pattern applied were imperative ones, which can legitimately act as sentences (6.8). Therefore instances of this rule application are covered by the NP=NP:S rule.

  (6.8) Meanwhile stations are fuming because many of them say, the show's distributor, Viacom Inc, is giving an ultimatum: either sign new long-term commitments to buy future episodes or risk losing "Cosby" to a competitor.

- VP=VP:NP — this rule-pattern is the result of a case of misbracketing (6.9). The colon-expansion should not be bracketed as an adjunct to the VP but rather as an adjunct to the whole sentence in order to make linguistic sense.

  (6.9) The following were neither barred nor suspended: Stephanie Veselich Enright, Rolling Hills, Calif., fined $2,500 and ordered to disgorge $11,762; Stuart Lane Russel, Glendale, Calif., fined $2,500 and ordered to disgorge $14,821; Devon Nilson Dahl, Fountain Valley, Calif., fined $82,389.

It should be noted, however, that whilst all the twelve patterns in table 6.2 are valid, not all of them are normal colon expansions. There are seven exceptions. Significantly though, all the rule-patterns are in agreement with the description of colon use that can be found in publishers' style guides ((Jarvie, 1992), for example), which even cite the exceptional cases found here.

- PP=P:NP — this rule-pattern uses the colon merely to introduce a conjunctive structure (6.10) — possibly one which is structurally separated from the preceding sentence fragment in, say, an itemised list and that has quite linguistically complex items.

  (6.10) We like climbing up: rock, trees and cliffs

- VP=V:NP & VP=V:S — these rule-patterns are similarly used to introduce conjunctive lists where the verb subcategorises for sentences or noun phrases, and also in certain writing styles to introduce direct speech (6.11).

  (6.11) They said: "We went to the party."

- NP=NP:PP — the only instance in the whole corpus of this rule-pattern was a book title (6.12). It unlikely to be used more frequently in any other circumstances.

(6.12) "Big Red Confidential: Inside Nebraska Football"

- PP=PP:PP — this is possibly the most potentially productive of the excepted rules. This pattern provides only for a colon expansion containing a clarifying PP re-using the same preposition (6.13). Its use is very infrequent, though.

(6.13) [...] spoke specifically of a third way: of having produced a historic synthesis of socialism and capitalism.

- S=PP:S — this rule-pattern is an exception since the mother category is not really a sentence (6.14). It is more likely to be an item in a list that is introduced by a phrase such as "*Views were aired on the following matters:*". The frequency of this pattern in the corpus is an artifact of its journalistic nature.

(6.14) On China's turmoil: "It is a very unhappy scene," he said.

- S=VPING:NP — this is a unique rule pattern whose mother category is not strictly speaking a grammatical sentence (6.15). There are two solutions — the initial verbal phrase can be treated either as a sentence with a null subject or as a gerund noun-phrase.

(6.15) Also spurring the move to cloth: diaper covers with velcro fasteners that eliminate the need for safety pins.

By repeating this pattern elimination for all the rules, the number of rule patterns were reduced to just 79, and more than half of these related to the comma. The rules are shown in table 6.3. Since some of the patterns only apply in particular, exceptional cases, the number

NP=NP:NP	NP=NP:ADJP	PP=PP:PP	VP=V:NP	S=S:NP	S=PP:S
NP=NP:S	NP=NP:PP	PP=P:NP	VP=V:S	S=S:S	S=VPING:NP
NP=NP;NP	S=S;S	VP=VP;VP	PP=PP;PP		S=PP.
S=INTJ.	S=S.	S=ADJP.	S=ADVP.	S=NP.	S=VP.
VP=VP-VP-	PP=PP-PP-	NP=NP-NP-	NP=NP-VP-	NP=NP-S-	NP=NP-PP-
S=S-S-	ADJP=ADJP-ADJP-	S=S-PP-	S=S-NP-		
ADJP=ADJP,	ADJP=ADJP,ADJP	ADJP=ADJP,ADVP	ADJP=ADJP,PP	ADJP=ADJP,S	
VP=VP,	VP=VP,VP	VP=VP,PP	VP=VP,S	VP=VP,NP	VP=VP,ADVP
ADVP=ADVP,	ADVP=ADVP,ADVP	ADVP=ADVP,SBAR		VP=ADVP,VP	VP=VP,ADJP
NP=NP,	NP=NP,NP	NP=NP,S	NP=NP,VP	NP=NP,PP	NP=NP,ADJP
NP=NP,ADVP	NP=ADVP,NP	NP=INTJ,NP	NP=PP,NP	NP=ADJP,NP	NP=VP,NP
S=S,	S=S,S	S=S,NP	S=S,VP	S=S,PP	S=S,ADVP
S=S,INTJ	S=INTJ,S	S=ADVP,S	S=PP,S	S=NP,S	S=VP,S
S=CONJ,S		PP=PP,	PP=PP,PP	PP=PP,ADVP	PP=ADVP,PP

Table 6.3: Processed underlying punctuation rule patterns

of 'standard' rules is reduced even further. Also, since many valid rule-patterns occur infrequently in the corpus, there exists the possibility that there are further valid infrequent punctuation patterns that do not occur in the corpus. Whilst some of these may be hypothesized, and incorporated into a formalisation, other more obscure patterns may be missed, and so the guidelines postulated here are not necessarily exhaustive for the whole language (especially since the corpus is from a single American english source). They should be sufficient, however, to characterise most of the usages of the punctuation marks that are likely to be encountered.

Formalism and Generalisation

If the exceptional cases are ignored, it is relatively straightforward to postulate some generalisations about the use of the various punctuation marks.

Colon expansions seem only to occur in descriptive contexts. Thus their mother category can be either NP or S, descriptive categories, rather than the active VP or locative PP. The mother category of a colon expansion is always the same as the category to which the adjunct is attached (the left-most daughter) and this is even true of many of the exceptional rule patterns if the constraint is relaxed to allow the daughter to have a lower bar-level. The phrase contained within the colon-expansion (right-most daughter) must also be descriptive, but can be ADJP in addition to NP and S. (Although there was no rule pattern found in the corpus that had an adjectival colon expansion with a sentential mother-category, it is certainly possible to imagine such a sentence (6.16).) Therefore (6.17) can be postulated as a general colon-expansion rule.

(6.16) The cat lay there happily: relaxed and warm.

(6.17) $\mathcal{X} = \mathcal{X} : \{ \text{NP} \mid \text{S} \mid \text{ADJP} \}$ $\mathcal{X}{:}\{\text{NP}, \text{S}\}$

The rule generalisation for semicolons is very simple, since the semicolon only separates similar items (6.18). The possibility exists that this rule may apply to further categories such as adjectival and adverbial, although instances of this were not found in the corpus.

(6.18) $\mathcal{S} = \mathcal{S} \; ; \mathcal{S}$ $\mathcal{S}{:}\{\text{NP}, \text{S}, \text{VP}, \text{PP}\}$

The generalisation for the full-stop is also straightforward, since it applies to all categories. The only problem is that it is not necessarily suitable for all the resulting structures to be referred to as sentences. The mothers should really all be top-category, since the full-stop is used to signal the end of a text-unit. Thus the generalisation in (6.19) is the most appropriate.

(6.19) $\mathbf{T} = * \; .$

The dash interpolation is the first punctuation mark for which generalisation becomes slightly complicated. There appear to be two general rules, which overlap slightly. The first (6.20) simply states that a dash interpolation can contain an identical category to the phrase it follows. The second rule (6.21) extends this rule when applied to the two descriptive categories, so that a wider range of categories are permitted within the interpolation — again, one of the rule-patterns permitted by (6.21) does not actually occur in the corpus, but does seem plausible.

Note that since these rules incorporate a final dash, they will rely on Nunberg's (1990) principle of point absorption to delete the final dash if necessary. In fact, on examination of the two rules, it becomes apparent that there is an area of overlap between them: those cases where the category within the dash interpolation is either sentential or nominal. Therefore the pair of rules is better stated as in (6.22).

(6.20) $\mathcal{D} = \mathcal{D} - \mathcal{D} -$ $\mathcal{D}$:{NP, S, VP, PP, ADJP}

(6.21) $\mathcal{E} = \mathcal{E} - \{$ NP | S | VP | PP $\} -$ $\mathcal{E}$:{NP, S }

(6.22) a. $\mathcal{D} = \mathcal{D} - \mathcal{D} -$ $\mathcal{D}$:{VP, PP, ADJP}

b. $\mathcal{E} = \mathcal{E} - \{$ NP | S | VP | PP $\} -$ $\mathcal{E}$:{NP, S }

The commas have the most complicated set of rule-patterns. The generalisation seems to be that any combination of phrasal categories is acceptable, so long as one of the daughter categories is identical to the mother category (6.23). The restriction on this, and the reason why there are fewer rule-patterns for categories such as PP, ADJP and ADVP, is that rules with the same daughters but more 'powerful' mother categories (e.g. sentential vs. adverbial) seem to be able to block the application of the 'less powerful' rules.

(6.23) a. $\mathcal{C} = \mathcal{C}$, $*$ $\mathcal{C}$:{NP, S, VP, PP, ADJP, ADVP}

b. $\mathcal{C} = *$, $\mathcal{C}$ $\mathcal{C}$:{NP, S, VP, PP, ADJP, ADVP}

As an extension to these results of the analysis, it is relatively straight-forward to postulate the following simple rules (6.24–6.27), even though the punctuation symbols they refer to are not explicitly searched for in this analysis, and they can in fact be verified in corpora.

- For any sort of quotation-marks (excluding so-called "Victorian Quotation"). Note also that Nunberg's principle of quote-transposition is captured almost by default by this rule, since any punctuation feature present inside the final quotation mark will be presented to the surrounding parse too.

  (6.24) $\mathcal{Q} =$ " $\mathcal{Q}$ " $\mathcal{Q} : *$

- For the various stress-markers

  (6.25) $\mathcal{Z} = \mathcal{Z}$? $\mathcal{Z} : *$

  (6.26) $\mathcal{Y} = \mathcal{Y}$! $\mathcal{Y} : *$

  (6.27) $\mathcal{W} = \mathcal{W}$ ... $\mathcal{W} : *$

This set of Generalised Punctuation Rules, then, seems to encapsulate the contextual positioning possibilities for punctuation marks among syntactic categories, and should not be too problematical to implement in most grammatical formalisms. However, in their current form it seems very likely that the use of these rules would lead to overgeneration in any grammatical analysis system, since there is nothing to restrain the positions and manner in which they operate. Therefore the second area of investigation into syntactic punctuation function is necessary.

6.2 Theoretical Approach

The trouble with the preceding observational approach to syntactic punctuation functionality is that it does not capture the *syntactico-structural* significance of punctuation. The generalised punctuation rules that were derived from the experimental data describe the positioning of the various punctuation marks quite adequately, with respect to the surrounding syntactic categories and the dominating nodal category of the parse, but they do little to describe the overall structural significance of the punctuation or describe the syntactic situations in which the use of these generalised rules is valid. Therefore it seems desirable to approach the problem from the theoretical dimension too, and then to integrate the results of both approaches.

Conjoining Punctuation

At the most basic level, based both on Nunberg's (1990) work and on simple observation of text, there are only two broad categories of punctuation: the sort that comes between lexical items (in a conjoined list, for example), and the sort that goes around them. Dealing with the occurrence of the first category is relatively simple, since invariably the marks occur between syntactically-similar categories (6.28).

(6.28) a. I'd like sausage, fish, chips and vinegar.

b. I like to run, to skip, and to jump.

c. The dark, green, whispering trees seemed to beckon.

d. He was a distinguished academic geographer; he had wide-ranging research skills; he published books in all sorts of subject areas; he was also a devoted husband and father (Jarvie, 1992).

e. You stand over there — and I will stay here.

The marks that can occur in this category of punctuation function are the comma and the semi-colon. The dash can also, occasionally, appear but tends to do so rather infrequently. Most of the lists in (6.28) have a lexical conjunction present before the final list item, but it is clear that this is not mandatory.

The best explanation for the function of punctuation in this category is that it is performing a **coordinating** role. The precise nature of the coordination is picked up from the lexical coordination present in the list. Thus the punctuation marks are, in effect, functioning as anaphors (cataphors, actually, since their referent follows them). This is not really a syntactic function *per se*, rather one of a later, anaphoric resolution stage. However, from the point of view of syntactic analysis, the punctuation marks provide a connective function between the elements to either side of them. Thus the best interpretation of a comma-conjoined list, e.g. (6.28a), is as a flat structure, as exemplified in (6.29). Several points immediately emerge from this, such as that punctuation marks adjacent to the actual lexical coordination, as in (6.28b), do not pick up the coordinating function, since this would result in a functional repetition (6.30). Also there

are cases where there is no lexical coordination to reference (6.28c); in these cases, the punctuation marks seem to function in a manner analogous to a lexical conjunct, at least syntactically (6.31).

(6.29) I'd like sausage and fish and chips and vinegar

(6.30) *I like to run and to skip and and to jump.

(6.31) The dark (and) green (and) whispering trees …

The reformation of the punctuation marks into a syntactically flat structure (as in (6.29) suggests another function of this category of punctuation mark, namely to 'chunk' the text appropriately. The punctuation marks seem to form a hierarchy, both with respect to one another and with respect to lexical coordinations, as illustrated in (6.32). This enables complex coordinated structures to be correctly aggregated, so that a properly representative syntactic structure will emerge if there is more than one sort of coordinating device present (6.33).

(6.32) **;** ≫ **—** ≫ **,** ≫ **and** ≫ **&**

(6.33) a. We sell egg and chips, sausage and chips, fish and chips, and spam and chips.

b. We sell egg and chips, at 80 pence; sausage and chips, at 70 pence; fish and chips, at 90 pence; and spam and chips, at 50 pence.

The dash as a coordinating punctuation mark is not only relatively infrequently used, but also seems only to be used in a restricted set of circumstances. Sentence (6.34) illustrates the use of the conjunctive dash: it can only be used to coordinate two items (any more, as in (6.35), and the potential arises for confusion with a dash-interpolation) and ideally should be followed with a lexical coordination although this is not always necessary (6.36). If a lexical coordination is not present, however, the potential for confusion with a dash-interpolation again arises. Of course, by the notions developed earlier, if the dash is followed by a lexical coordination, it does not referentially adopt the function of that coordination, since that would result in a functional repetition. Since only one coordinating dash can occur in an item, however, this means that the sole function of the dash in the presence of a lexical coordination is to provide correct aggregation, as in (6.37) where it separates the preceding, lower-level conjunction from the following item in the contrastive coordination. In addition, the items coordinated by the dash must be of syntactic level equivalent to a sentence: if the items were of any lower level, as in (6.38), the dash-interpolation reading is obtained.

(6.34) I enjoy shopping at Tesco's — but you prefer Sainsbury's.

(6.35) ~I enjoy shopping at Tesco's — but you prefer Sainsbury's — and Gillian uses Safeway.

(6.36) ~I enjoy shopping at Tesco's — you prefer Sainsbury's.

(6.37) ~I enjoy shopping at Tesco's and Gillian uses Safeway — but you prefer Sainsbury's.

(6.38) ~I am going to Tesco's — and Sainsbury's.

It should be noted that as a corollary of the cataphoric inheritance of coordinative functionality by the punctuation marks, if two or more *different* lexical coordinations are present in the list, this causes the usual linear analysis to fail, and a different structure should be attempted (if possible). Thus while the multiple conjuncts in (6.39) provide no problem, the conjunct and disjunct in (6.40) clash and force a questionable analysis that perhaps the first two items are conjoined, and the resulting entity is then disjoined from the final item.

(6.39) I like fish, and chips, and sausages.

(6.40) ?I want apples, and pears, or bananas.

Thus the principles to emerge from this category of punctuation are roughly as follows (with obvious featural restrictions to prevent attachment ambiguities in lists of length greater than two, for example to force right branching in analyses). Note the slight difference in the formulation of the rule for the dash to prevent lists longer than two items.

$\mathcal{X} \Rightarrow \mathcal{X}$ **, (coord)** $\mathcal{X}$	$\mathcal{X} : *$
$\mathcal{Y} \Rightarrow \mathcal{Y}$ **; (coord)** $\mathcal{Y}$	$\mathcal{Y} : *$
$\mathcal{Z}' \Rightarrow \mathcal{Z}$ **— (coord)** $\mathcal{Z}$	$\mathcal{Z} : *$

Adjoining Punctuation

The operation of punctuation in coordination, then, is relatively straightforward. More problems arise with the use and operation of the non-coordinative, more syntactically-contentful punctuation marks. One role that punctuation marks seem to perform is to mark phrasal boundaries. Therefore a promising starting point for the theoretical investigation of these marks is to propose that they can occur occur between phrasal-level (or higher) items (e.g. NP, VP, PP, ADVP, S).

(6.41) Those that can, contribute to the fund. S[NP,VP]

(6.42) Recently, I went out. S[ADVP,S]

(6.43) The man, my friend, is here. S[NP(NP,NP),VP]

(6.44) The man, with the stick, is here. S[NP(NP,PP),VP]

(6.45) *Rich republicans, contribute to the fund. S[NP,VP]

(6.46) His, but not her, dog won the contest. [NP(DET(DET,DET),N)VP]

Sentences (6.41) and (6.42) clearly conform to this principle, with the single punctuation mark occurring between phrasal or higher-level items. The principle can also be seen operating at multiple levels to explain examples of Nunberg's (1990) delimiting punctuation too, in (6.43) and (6.44). However, (6.45) has a superficially similar structure to (6.41), and yet the punctuation placement in this sentence is infelicitous. Contrastingly, the punctuation use in (6.46) is felicitous but not licensed by the principle hypothesized, since neither the noun, nor

the complex determiner, are at a phrasal level (although an argument could be made for the presence of a 'determiner-phrase' if such a constituent were to exist).

The hypothetical principle therefore clearly does not operate as stated. Not only does the principle as stated block appropriate punctuation uses, but it also permits infelicitous ones. Additionally, the definition of 'phrasal level or above' is a rather ill-defined and untidy one, since under X-bar theory (Jackendoff, 1977) some of the syntactic categories referred to are at bar-level 1 (verb phrase), and others are at bar-level 2 (the others).

Some notion of adjacency of punctuation to phrasal level entities seems desirable, however, since it many cases punctuation does seem to mark phrasal boundaries. Therefore it would be possible to define 'phrasal' as referring to any complex linguistic structure, so that punctuation could occur between any such complex linguistic structures. This would, however, again permit (6.45), and block (6.46) since the noun *dog* is not phrasal. In order to admit (6.46), the principle could be redefined so that instead of forcing punctuation to occur between two complex items, it only need occur adjacent to one such item. Whilst this reformulation now permits (6.46), it also still permits (6.45). In addition, it would be necessary to prevent occurrences such as (6.47), which would also be licensed under this present rule formulation using one-sided adjacency. Using a different definition of the term 'phrasal' then, punctuation could be licensed only when adjacent to maximal level phrases (e.g. NP, S). These are the phrases that under X-bar theory (Jackendoff, 1977) would be described as level 2. However, this rules out felicitous cases like (6.48), whilst still permitting (6.45).

(6.47) The, new toy …

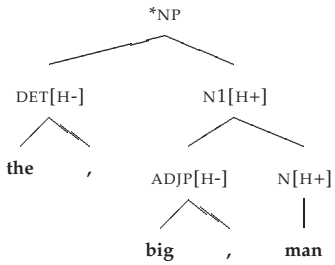
(6.48) He does, surprisingly, like fish.

Clearly, therefore, something stricter than the first approach, but more relaxed than the second, is needed. Instead of the rather informal notions of phrasal entities and attachment thereto, a neater and more rigorous approach would be to use the notion of headedness. Use of such an explicitly lexical-syntactic phenomenon for the processing of punctuation is yet another argument for the full integration of punctuation and Nunberg's text grammar into the lexical grammar.

Using head features, the postulation of a new version of the principle so that only non-head structures can have punctuation attached still does not rule out examples like (6.49), since the head-feature is still absent in all the syntactic items that the comma is attached to (6.50). This is also a departure from the previous formulation in that punctuation marks are considered to be attached to syntactic entities rather than floating freely between them. Further refinement is necessary, and the answer seems to be related the bar-levels of daughter and mother categories under X-bar theory. Attachment of punctuation to the non-head daughter only seems to be legal when mother and head-daughter are of the same bar level, regardless of what that bar level is (and indeed usually when they are identical categories, since it is very unusual to have a rule where the head-daughter and mother are different syntactic categories if they have the same bar level).

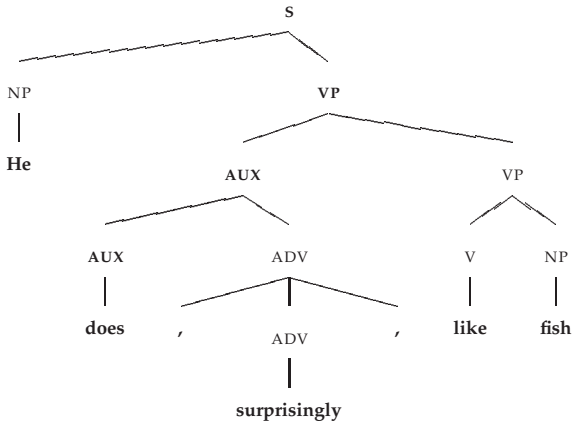
(6.49) *the, big, man

(6.50)



This version of the principle seems to work: (6.47) and (6.45) are blocked, while the felicitous (6.48) and (6.46) are sanctioned. The sanctioned interpretation of (6.48) is as shown in (6.51), where the main head-feature progression is shown by syntactic categories in **bold** font. As an aside, it becomes obvious in examples such as (6.48), that the apparent attachment of the first comma to the preceding item is a purely orthographical convention. In linguistic structure, as shown in (6.51) the punctuation marks both attach to the non-head structure, on either side of it.

(6.51)



The theoretical principle formulated above can now be restated more neatly as follows, where the term *structure* refers either to a rule application (in a grammatical context) or a node (in the context of a representational structure).

> **Punctuation marks can be syntactically attached to any item provided the resulting syntactic entity functions as a non-head daughter in a structure where the head daughter and the mother categories have the same bar level under X-bar theory.**

From this theoretical approach it appears that this category of punctuation marks could be described as being adjunctive (i.e. those phrases to which punctuation is attached perform an adjunctive function). Since these adjunctive phrases can occur both at the extreme left (6.52) and right (6.53) of a sentence, there are not always two punctuation marks surrounding them, as in (6.54). By taking heed of the principle of point absorption (Nunberg, 1990), we can declare that the natural form of all such adjunctive structures is to be delimited with a punctuation mark at both extremities. It is only when these punctuation marks overlap with others that certain absorption phenomena take place to remove one of them.

(6.52) Additionally, we have been entered for the big prize.

(6.53) I bade farewell to my best friend, Arthur Smith.

(6.54) Arthur Smith, my best friend, is coming to stay.

Thus there is a similarity between these adjunctive punctuation marks and Nunberg's concept of *delimiting* punctuation. However, adjunctive punctuation, as defined here, is far richer than the delimiting marks in (Nunberg, 1990). Classifying the marks in terms of their similar syntactic application, the adjunctive marks include paired commas, paired dashes, paired brackets, the combination of the sentence-initial capitals and the full-stop, and also the colon-expansion, where the colon is paired with a null marker that is always either absorbed by the end of the sentence or a semi-colon. The reason that the colon-expansion is categorised with adjunctive punctuation, despite the lack of a matching final mark, is that the content of a colon-expansion is adjunctive to the material before.

Thus with this richer set of punctuation marks the hierarchy dictating point absorption (3.53) must be modified slightly to pair the sentence-initial capitalisation phenomenon with the full-stop in terms of power. The reason the full-stop is treated as adjunctive rather than conjunctive is that it is manifestly paired with the sentence-initial capitalisation, which in turn is also manifestly a punctuation phenomenon. Therefore, since the phenomena are paired, and since the conjunctive contribution of a sentence-break is questionable anyway (discussed further in chapter 8), the adjunctive interpretation is the most suitable.

Depending on the head-feature conventions used in formulating conjunction principles in a grammar, conjunctive uses of punctuation (6.55) could also be treated under the adjunctive principle. (In fact we can do this in all cases involving binary conjunction except that where both constituents of the conjunction are treated as carrying the head feature.) This is possible since the punctuation can always be attached to that element of each binary conjunction which does not carry the head feature. Since the result of each conjunction is an identical syntactic category to the conjoined items (which therefore has the same bar level) the conditions have been achieved for the principle to be able to apply. The only difference in this case between the two categories of punctuation is that conjunctive punctuation is always singular, whereas adjunctive punctuation is always paired (at least at a deep, if not a surface, level). However, the function of conjunctive punctuation is not an adjunctive one, so this merging of the two types, while an interesting aside, is not suitable for the realisation of the function of the punctuation marks involved.

(6.55) dogs, cats, fish and mice

(6.56) He said, "Welcome to my humble home."

There are a few other phenomena of punctuation that have not been covered so far. Quotation marks, at a surface level, approximate the adjunctive punctuation category. However, the classification of a quoted phrase as an adjunct is totally dependent on the subcategorisation of the introducing verb. In (6.56), if the subcategorisation of *said* is sentential, the quoted phrase cannot be adjunctive. If, however, the subcategorisation is empty, then the adjunctive function can apply. The situation is made easier by the observation that quoted material is often separated from the surrounding non-quoted material by means of commas. The function of these commas in this case cannot be conjunctive, and so must be adjunctive. Therefore an empty subcategorisation is required anyway, and the quotation marks can be adjunctive in function. (Of course, here there is then the further problem of whether to treat the lower-level adjunct as the head of the higher-level one, in which case the principle fails…) This does not matter, anyway, since quotation operates somewhat differently to other punctuation anyway, in that it does not appear to have to signal any phrasal boundaries at all, as illustrated in (6.57).

(6.57) According to FBI agent David ("Woody") Johnson, "a white male with an indistinguishable" American accent warned that a bomb would go off at the park within 30 minutes. (Gleick, 1996)

Thus the best way to treat quotation marks is to regard them as superficially adjunctive, but syntactically non-functional (transparent to syntactic analysis). This is confirmed in Doran's subsequent analysis of quotation (Doran, 1996) where she shows that quotation marks are not adequate to either identify or constrain the text fragments they surround. Their function seems to be at a semantic level, marking the enclosed text as being originally generated by another 'author'.

Stress markers are also punctuational devices that should be treated as syntactically transparent, except, of course, in those circumstances where they have absorbed a full-stop, in which case they take on the functionality of that adjunctive punctuation mark.

The final problem of the theoretical principles hypothesised above is that they appear to fail to account for the original sanctioned example of the chapter, repeated as (6.58). This usage of the comma, denoted by Nunberg (1990) as *disambiguating*, distinct from his delimiting and separating uses, appears not to be sanctioned since it cannot be conjunctive and appears singularly in the middle of a sentence, which appears to be invalid adjunctively. However, it is possible to treat this mark as adjunctive, when we consider the nature of the noun phrase that precedes it, and contrast it with (6.59).

(6.58) Those who can, contribute to the fund.

(6.59) *Rich Republicans, contribute to the fund.

In the case where the use of the comma is sanctioned, the noun phrase that precedes it is a complex one, containing the adjunctive clause *who can*. If the theoretical principle for

adjunctive punctuation permits two marks to surround an adjunct, it also permits one to occur. Therefore the disambiguating use of the comma is a special case of the adjunctive punctuation principle in which only a single punctuation mark occurs, even at the deep level of representation. This therefore allows the principle controlling the application of adjunctive punctuation to embody within it the notions associated with Nunberg's point-absorption principle, since the adjunctive punctuation marks are not forced to be paired. Thus if, for example, a final adjunctive comma is absorbed by a semi-colon then the adjunctive principle can still operate with the single initial adjunctive comma, before allowing the resultant syntactic entity to be passed on to be conjoined via the semi-colon.

Therefore, working theoretically, the class of inter-lexical punctuation has been split into two categories: conjunctive and adjunctive (or possibly coordinating and subordinating). Principles have also been presented for the syntactic operation of these categories and to ensure their coverage of punctuation phenomena. It remains to see how the theoretical work will integrate with the observational results, and whether the combined results can be applied profitably in a computational context.

6.3 Other Related Work

Subsequently and independantly to the work described here, some other principled approaches to including punctuation in computational systems have been made. Briscoe (1994) bases his treatment on Nunberg's theory of punctuation, with a grammar of 26 rules that are claimed to capture most of Nunberg's text-sentential constraints. These rules can be used on their own to indicate the rough structure of a sentence, as indicated by the punctuation marks it contains (and in which case the lexical items occurring between punctuation marks are treated uniformly and grouped into a flat textual unit) and are also integrated to a certain extent with a proper syntactic part-of-speech tag based grammar.

Briscoe agrees that it is undesirable to separate the punctuation processing and parsing stages of analysis due to the potentially high level of ambiguity in analysis of just punctuation, and the possibility of resolving many of the ambiguities through an at least partial integration of syntactic and punctuation analysis.

Unlike the early approach described in chapter 4, Briscoe retains the modularity of the punctuation grammar by folding the lexical grammar into the textual one, and dealing with the properties of the two grammars using disjoint sets of features. The text grammar rules are represented as left or right branching rules of 'Chomsky-adjunction' to lexical and phrasal rules. The most significant difference from the representation discussed in chapter 4, however, is that Briscoe choses to treat punctuation marks as seperately tokenised entities, whereas the approach taken earlier effectively treats them as clitics on the words they follow, at a cost of one extra feature per lexical item.

Briscoe does encounter some problems with this approach, in particular with certain uses of the comma, and so has to treat these cases in a more integrated manner to the other punctuation, implementing the inclusion of punctuation in more normal, low-level, grammatical rules.

Doran, in her investigation of punctuation in quoted speech (1996), approaches the problem

in a similar way. Based on the theories of Nunberg, she only treats quotation marks and punctuation associated specifically with quotation phenomena (colons, commas, etc). This makes her account somewhat ad-hoc from the point of view of general punctuation analysis, but still theoretically well-motivated within the domain of investigation. Doran's investigations are carried out using LTAG's, however, which lends support to the generalisability of this sort of approach across different language analysis platforms.

seven Testing the Syntactic Notions of Punctuation

And now here I was in Illinois, and it was flat and full of corn and boring. A childlike voice in my head cried, 'When are we going to be there? I'm bored. Let's go home. When are we going to be there?' Having confidently expected at this time to be in Missouri, I had my book of maps opened to the Missouri page, so I pulled over to the side of the road, in a state of some petulance, to make a cartographical adjustment. A sign just ahead of me said BUCKLE UP. ITS THE LAW IN ILLINOIS. *Clearly, however, it was not an offence to be unable to punctuate.* (Bryson, 1990)

Reading on from the observational and theoretical explorations of syntactic punctuation function in the previous chapter, it is now necessary to try to combine the two sets of results to see if they are complementary, and to test whether or not the Generalised Punctuation Rules and theoretical principles arrived at are indeed valid and useful. Therefore, the best course of action is to incorporate them into a grammar and use it in a parsing exercise. Once again, as in (Jones, 1994b) and (Briscoe and Carroll, 1995), the most flexible and functional parsing framework to use will be the Alvey Tools' Grammar Development Environment (GDE) (Carroll et al., 1991), which allows for rapid prototyping and easy analysis of parses.

The grammar used as a basis for this study is a tag-based one developed from that used in (Jones, 1994b) (which, in turn, was developed from the one used in (Briscoe and Waegner, 1992)). The corpus used is a version of the SUSANNE corpus (Sampson, 1995), prepared and retagged by John Carroll. This corpus consists of just over 7,000 sentences from the Brown corpus, used in (Briscoe and Carroll, 1995).

As presented in the previous chapter, the theoretical principles for syntactic punctuation function can be neatly summarised via some rules and principles. The conjunctive or coordinating function of punctuation can be summarised via the rule schema shown in (7.1) and the

aggregational power-hierarchy shown in (7.2).

(7.1) $\mathcal{A} \Rightarrow (\mathcal{A}\mathcal{P})^{+}\ \mathcal{A}$ $\quad \mathcal{A} = *$ $\quad \mathcal{P} = \{ ; , — \}$

(7.2) ; ≫ — ≫ , ≫ **lexical** ≫ **&**

The situation is slightly more complicated with the adjunctive punctuation function, since the rule schemata should really incorporate some of the features of the punctuation absorption principles. Thus rather than one simple schema, it is necessary to have seven, as shown in (7.3). Additionally, there are restrictions on the way some of these rule schemata operate, for example relating to the possible scoping of the colon-expansion.

(7.3)
$\mathcal{B} \Rightarrow \mathcal{B}[\mathcal{H}+]\ \mathcal{P}\ \mathcal{X}[\mathcal{H}-]\ \mathcal{P}$ $\quad \mathcal{B} = *$ $\quad \mathcal{P} = \{ — , \}$ $\quad \mathcal{X} = *$
$\mathcal{B} \Rightarrow \mathcal{B}[\mathcal{H}+]\ (\ \mathcal{X}[\mathcal{H}-]\ )$ $\quad \mathcal{B} = *$ $\quad \mathcal{X} = *$
$\mathcal{B} \Rightarrow \mathcal{B}[\mathcal{H}+]\ \mathcal{P}\ \mathcal{X}[\mathcal{H}-]$ $\quad \mathcal{B} = *$ $\quad \mathcal{P} = \{ : — , \}$ $\quad \mathcal{X} = *$
$\mathcal{B} \Rightarrow \mathcal{B}[\mathcal{H}+]\ \mathcal{X}[\mathcal{H}-]\ \mathcal{P}$ $\quad \mathcal{B} = *$ $\quad \mathcal{P} = \{ , \}$ $\quad \mathcal{X} = *$ (special case)
$\mathcal{B} \Rightarrow \mathcal{P}\ \mathcal{X}[\mathcal{H}-]\ \mathcal{P}\ \mathcal{B}[\mathcal{H}+]$ $\quad \mathcal{B} = *$ $\quad \mathcal{P} = \{ — , \}$ $\quad \mathcal{X} = *$
$\mathcal{B} \Rightarrow (\mathcal{X}[\mathcal{H}-]\ )\ \mathcal{B}[\mathcal{H}+]$ $\quad \mathcal{B} = *$ $\quad \mathcal{X} = *$
$\mathcal{B} \Rightarrow \mathcal{X}[\mathcal{H}-]\ \mathcal{P}\ \mathcal{B}[\mathcal{H}+]$ $\quad \mathcal{B} = *$ $\quad \mathcal{P} = \{ , \}$ $\quad \mathcal{X} = *$

Thus while it would be interesting to see how good the results would be from the integration just of these schemata into a normal grammar, this is unlikely to work in practice, since extra information and features (as discussed in chapters 4 and 6) are needed. Additionally, it would have been very difficult to insert these rule schemata into the grammatical formalism being used in the current study, since the general categorial assignments (e.g. for the content of the adjunctively delimited phrases) would be very difficult to represent.

Thus restrictions are needed on the operation of the rule schemata above, particularly regarding the possible syntactic categories that can be used to instantiate the schemata, and in terms of multiple-rule ambiguity. The sentences (7.4) and (7.5), for example, could be ambiguously analysed with both the first and the fifth adjunctive rule schema in (7.3), to attach the adjunctive punctuation-delimited phrase to either the preceding NP or the following VP. To prevent such ambiguities, more concrete rules are needed to implement specific situations for punctuation usage, which although based on the theoretically-derived rule schemata will have a far narrower scope of operation.

(7.4) The man, who is my friend, saw the fire first.

(7.5) The Italians, apparently, are going to win the cup.

The most appropriate manner in which the theoretically-derived rule schemata can be constrained is with the generalised punctuation rules (GPR's) that were extracted from a parsed corpus in the observational investigation in chapter 6. As it stands, the theoretical principles of the previous chapter integrate very well with the GPR's extracted from the corpora. All the GPR's have the potential to comply in parsing use with the restrictions and constraints made

by the theoretical principles, and conversely the GPR's provide a valuable constraint to the possible scope and applicability of the theoretical principles. Neither set of principles would be successfully used on its own: as discussed above, the direct implementation of the rule schemata would lead to overgeneration and ambiguity, but similarly the GPR's do not provide enough structural detail or syntactic context for sole successful operation. The comma GPR's, for example (7.6), give no account of how the adjunctive usage of the comma functions to delimit an item on both sides, unless some absorption phenomenon takes place.

(7.6) $\mathcal{C} = \mathcal{C}$, $*$ $\mathcal{C}$:{NP, S, VP, PP, ADJP, ADVP}
$\mathcal{C} = *$, $\mathcal{C}$ $\mathcal{C}$:{NP, S, VP, PP, ADJP, ADVP}

It is therefore clear that the two complementary accounts of punctuation function in syntax, the theoretical and the observational, must be combined to give an implementationally tractable and concise description of the syntactic function of punctuation marks. The generalised punctuation rules shown in the previous chapter were converted into phrase structure rules, via the theoretical priciples relating to the adjunctive nature of phrases to which punctuation is attached. The resulting set of rules include extra functionality to embody the sense of the relevant parts of Nunberg's (1990) text-grammar (such as nesting of text adjuncts) and his absorption principles. The resulting grammar is not necessarily the neatest or most efficient, but it will serve to illustrate whether the principles of syntactic punctuation function derived in the previous chapter are applicable.

The colon rule was implemented by the creation of a new category for the various colon-expansions (7.7), and two rules to subsequently attach these adjunctive expansions to a head (7.8).

(7.7) a. ColExp $\dashrightarrow$ [punc colon] N2.
b. ColExp $\dashrightarrow$ [punc colon] V2.
c. ColExp $\dashrightarrow$ [punc colon] A2.

(7.8) a. V2[] $\longrightarrow$ H2[] ColExp.
b. N2[WH @x] $\dashrightarrow$ H2[WH @x] ColExp.

The semicolon rule was implemented by five rules of the form shown in (7.9). To prevent extra ambiguity in parsing due to various possibilities of attachment in sentences with two or more semi-colons, a feature was used to ensure right-branching (as in (Jones, 1994b)).

(7.9) V2[sc +] $\longrightarrow$ H2[sc -] [punc semi] V2[CONJ @x].

The two dash rules are dealt with separately. The initial one is implemented by straightforward rules (7.10) whilst the second one, similar to the case with the colon, requires a new category to be created for dash-interpolations (7.11) and rules to attach this category to other constituents (7.12). The mechanism that deals with point absorption in these rules is rather crude, but sufficient for the current experiment. Also a similar mechanism is used to prevent multiple attachment as is used in the semi-colon rules.

(7.10) a. V1[da +] $\longrightarrow$ H1[da -] [punc dash] V1[da -] ([punc dash]).
b. P2[da +] $\dashrightarrow$ H2[da -] [punc dash] P2[da -] ([punc dash]).
c. A2[da +] $\dashrightarrow$ H2[da -] [punc dash] A2[da -] ([punc dash]).

(7.11) a. DashInt $\longrightarrow$ [punc dash] N2[da -] ([punc dash]).
b. DashInt $\longrightarrow$ [punc dash] V2[da -] ([punc dash]).
c. DashInt $\longrightarrow$ [punc dash] V1[da -] ([punc dash]).
d. DashInt $\longrightarrow$ [punc dash] P2[da -] ([punc dash]).

(7.12) a. V2[da +] $\longrightarrow$ H2[da -] DashInt.
b. N2[da +] $\longrightarrow$ H2[da -] DashInt.

The comma rule is the most tricky to implement, and also seems the most likely candidate to generate extra ambiguity. Therefore, we will take notice of the result from observing the rule-patterns, namely that more powerful categories seldom attached as adjuncts to less powerful ones. Hence we can consider a hierarchy of categories, as in (7.13) and construct phrase structure rules accordingly.

(7.13) V2 $\gg$ N2 $\gg$ V1 $\gg$ P2 $\gg$ A2

Thus a head of type V2 can have any other category attached to itself as a comma adjunct (7.14), but a head of type A2 can only have a comma adjunct of type A2 (7.15). To prevent a huge amount of over-generation, the final comma here is mandatory (making the use of commas here strictly delimiting). Single comma adjunct attachment is also provided, but since this typically means attaching an adjunct at the beginning or end of a sentence, this is provided for directly at sentential level. Therefore this solution is not the neatest way of tackling the problem, but it is relatively simple and will suffice for the current purposes.

(7.14) a. V2 $\longrightarrow$ H2 [punc comma] A2 [punc comma].
b. V2 $\longrightarrow$ H2 [punc comma] P2 [punc comma].
c. V2 $\longrightarrow$ H2 [punc comma] V1 [punc comma].
d. V2 $\longrightarrow$ H2 [punc comma] N2 [punc comma].
e. V2 $\longrightarrow$ H2 [punc comma] V2 [punc comma].

(7.15) A2 $\dashrightarrow$ H2 [punc comma] A2 [punc comma].

Of course, these rules have only served to cover the syntactic functionality of the adjunctive punctuation marks. The functionality of the conjunctive marks, as presented in chapter 6, is easily included in the normal grammatical rules covering category conjunction, by modifying these to be compatible with the generalised conjunction rules presented in the theoretical investigation (with additional provisos to cover semi-colon promotion).

With the punctuation rules included, the grammar contains around 300 simple rules. The results of parsing the SUSANNE corpus are shown in table 7.1 and figure 7.1. Since 262 sentences

Parses per sentence	Number of sentences	Percentage of corpus
0	2,166	32.1%
1–5	1,291	19.1%
6-10	444	6.6%
11–50	803	11.9%
51-100	310	4.6%
101–500	561	8.3%
501–1,000	185	2.7%
1,001–5,000	334	4.9%
5,001–10,000	134	2.0%
10k–50k	177	2.6%
50k–100k	52	0.8%
100k +	257	3.8%
timeouts	38	0.5%
sentences	6,752	

Table 7.1: Parsing results for SUSANNE corpus

contained words or constructs that the grammar had not been designed to deal with (e.g. sentence (7.16)), the zero results for those sentences have been ignored. Note that the parse numbers shown here are not true numbers of unpacked parses, but just estimates of the number of passible parses in the parse forest. Code to perform this estimation was kindly supplied by John Carroll. Testing has shown that this technique tends to slightly over-estimate the number of parses, but not by a very large amount.

(7.16) (she will return this symbol to the mountain , as one pours seed back into the soil every spring ... or as ancient fertility cults demanded annual human sacrifice)
** Error, word ... not found

These results are encouraging, especially since the core grammar used in the present study is suboptimal, and is also only a grammar of grammatical tags, and thus excludes features such as subcategorisation. This suggests that this is a promising way forward, and that the Generalised Punctuation Rules and theoretical principles derived here are applicable and useful to the problem of identifying punctuation function and operation.

Some sample analyses from parsing of punctuation use in the SUSANNE corpus are presented below.

7.1 Sample Analyses

Adjunctive commas are handled appropriately, regardless of whether the comma-delimited item occurs at the start of the sentence (7.17), in the middle of the sentence (7.18) and (7.19) or at the end of the sentence (7.20). The most appropriate parses of these sentences are shown in figure 7.2.

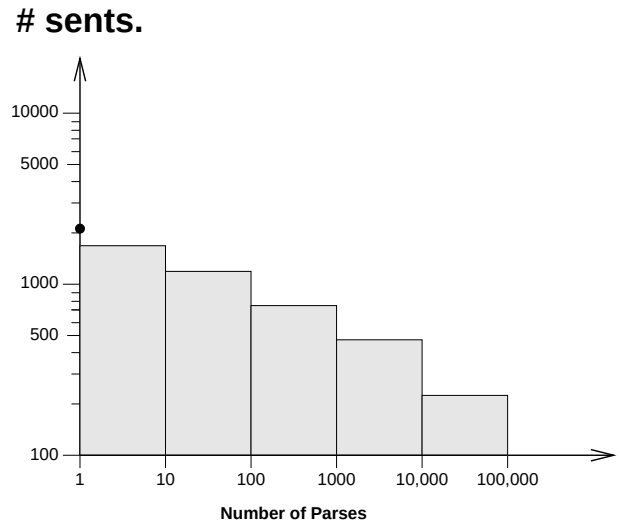


Figure 7.1: Parsing results for (SUSANNE) corpus (logarithmic axes)

(7.17) In_II some_DD areas_NN2 ,_, the_AT progress_NN1 is_VBZ slower_JJR than_CSN in_II others_NN2 *(6 parses)*

(7.18) His_APP$ light_NN1 blue_JJ eyes_NN2 ,_, set_VVD deep_RR within_II the_AT face_NN1 ,_, were_VBR actively_RR and_CC continually_RR looking_VVG *(15 parses)*

(7.19) The_AT creative_JJ urge_NN1 ,_, for_REX example_REX ,_, transcends_VVZ the_AT body_NN1 and_CC the_AT self_NN1 *(1 parse)*

(7.20) It_PPH1 is_VBZ infuriating_VVG ,_, this_DD1 feeling_NN1 that_CST one_PN1 is_VBZ being_VBG picked_VVN on_RP ,_, continually_RR ,_, constantly_RR *(21 parses)*

Of the other classic, delimiting adjunctive punctuation marks, only the dash was implemented in the grammar, and a sample sentence containing a final dash-interpolation is shown in (7.21). Of the possible parses of this sentence, the one presented in figure 7.3 seems the most appropriate. Other adjunctive punctuation marks, such as the colon, were analysed as in sentences (7.22,) illustrating usage to introduce speech, and (7.23), to illustrate the more linguistic adjunctive use. Appropriate analyses are shown in figure 7.4.

(7.21) He_PPHS1 steered_VVD me_PPIO1 to_II the_AT right_JJ track_NNL1 -_- the_AT free_JJ enterprise_NN1 track_NNL1 *(12 parses)*

(7.22) The_AT mayor_NNS1 said_VVD :_: it_PPH1 did_VDD not_XX come_VV0 from_II me_PPIO1 *(4 parses)*

```
((((in_II (some_DD (areas_NN2)))) ,_,
  ((the_AT (progress_NN1))
   (be_VBZ ((slower_JJR ((than_CSN ((in_II ((others_NN2)))))))))))
```

```
((((his_APP$ (light_NN1 (((blue_JJ)) (eyes_NN2)))) ,_,
   (set_VVD (((deep_RR)) (within_II (the_AT (face_NN1))))) ,_,)
  (were_VBR (((actively_RR))
             (and_CC ((((continually_RR)) looking_VVG)))))))
```

```
(((((the_AT (((creative_JJ)) (urge_NN1))) ,_, (for_REX example_REX) ,_,)
  (transcends_VVZ ((the_AT (body_NN1)) (and_CC (the_AT (self_NN1)))))))
```

```
(((it_PPH1 (is_VBZ (infuriating_VVG))) ,_,
  (this_DD1
   (feeling_NN1
    (that_CST
     (((one_PN1 ((is_VBZ (being_VBG (picked_VVN))) ((on_RP)))) ,_,
       ((continually_RR)))
      ,_, ((constantly_RR)))))))))
```

Figure 7.2: Appropriate parses of sentences (7.17–7.20)


```
((he_PPHS1
  ((steered_VVD me_PPIO1)
   ((to_II
     ((the_AT (((right_JJ)) (track_NNL1)))
      (-_- (the_AT (((free_JJ)) (enterprise_NN1 (track_NNL1))))))))))))
```

Figure 7.3: An appropriate parse of sentence (7.21)

```
((((the_AT (mayor_NNS1)) (said_VVD))
  (:_: (it_PPH1 (did_VDD (not_XX (come_VV0 ((from_II me_PPIO1)))))))))
```

```
((((I_PPIS1 (suppose_VV0 (the_AT (reason_NN1))))
   (is_VBZ (a_AT1 (kind_NN1)) ((of_IO (((wishful_JJ)) (thinking_NN1))))))
  (:_:
   (((do_VD0
      (not_XX
       (talk_VV0
        ((about_II
          (the_AT
           (((final_JJ))
            (stages_NN2 ((of_IO (reconstruction_NN1))))))))))))
    (and_CC
     (they_PPHS2
      (will_VM (take_VV0 (care_NN1 ((of_IO themselves_PPX2))))))))))
```

Figure 7.4: Appropriate parses of sentences (7.22) and (7.23)

(7.23) I_PPIS1 suppose_VV0 the_AT reason_NN1 is_VBZ a_AT1 kind_NN1 of_IO wishful_JJ thinking_NN1 :_: do_VD0 not_XX talk_VV0 about_II the_AT final_JJ stages_NN2 of_IO reconstruction_NN1 and_CC they_PPHS2 will_VM take_VV0 care_NN1 of_IO themselves_PPX2 *(47 parses)*

Of the conjunctive punctuation marks, only the semi-colon and the comma received a treatment in this grammar. Sentences that include semi-colons used to conjoin sentence-level elements are shown in (7.24) and (7.25). Similarly, conjunctive usage of commas is exemplified in sentences (7.26–7.29). In sentences (7.26) and (7.27), the commas reference to the final, lexical conjunction, which is also immediately preceded by a (redundant) comma. Sentence (7.28) illustrates a case of reference to a lexical disjunction, similarly preceded by a comma, and sentence (7.29) illustrates the usage not only of a comma referencing a lexical conjunction that is not preceded immediately by a comma, but also an exophoric conjoining comma linking the adjectives *antiseptic* and *windowless*. The four sentences all additionally include instances of adjunctive commas. The most appropriate structural analyses of the sentences involving the conjunctive semi-colons are presented in figure 7.5, and those of the four sentences involving the conjunctive comma are presented in figure 7.6.

(7.24) She_PPHS1 remained_VVD squatting_VVG on_II her_APP$ heels_NN2 all_DB the_AT time_NNT1 we_PPIS2 be_VBDR there_RL ;_; like_ICS the_AT man_NN1 ,_, she_PPHS1 was_VBDZ entirely_RR naked_JJ *(40 parses)*

(7.25) The_AT book_NN1 concerned_VVN with_IW the_AT Negro_NN1 's_$ role_NN1 in_II an_AT1 urban_JJ society_NNJ1 is_VBZ rare_JJ indeed_RGA ;_; recently_RR only_RR Keith_NP1 Wheeler_NP1 's_$ novel_NN1 ,_, Peaceable_JJ Lane_NNL1 ,_, has_VHZ openly_RR faced_VVD the_AT problem_NN1 *(80 parses)*

(7.26) In_II the_AT physical_JJ sciences_NN2 ,_, these_DD2 achievements_NN2 concern_VV0 electricity_NN1 ,_, chemistry_NN1 ,_, and_CC atomic_JJ physics_NN1 *(7 parses)*

(7.27) At_II Jenks_NP1 Street_NNL1 ,_, Simms_NP1 said_VVD ,_, the_AT car_NN1 skidded_VVD completely_RR around_RL ,_, just_RR missed_VVN two_MC park_VVN cars_NN2 ,_, and_CC sped_VVD East_ND1 in_II Jenks_NP1 *(260 parses)*

(7.28) If_CS one_MC1 dancer_NN1 slaps_VVZ another_DD1 ,_, the_AT victim_NN1 may_VM do_VD0 a_AT1 pirouette_NN1 ,_, sit_VV0 down_RP ,_, or_CC offer_VV0 his_APP$ assailant_NN1 a_AT1 fork_NN1 and_CC spoon_NN1 *(352 parses)*

(7.29) To_II the_AT men_NN2 in_II the_AT instrument_NN1 jamming_VVN bomber_NN1 cockpit_NN2 ,_, submarine_JJ compartment_NN2 and_CC the_AT antiseptic_JJ ,_, windowless_JJ room_NN2 that_CST will_VM be_VB0 the_AT foxhole_NN2 of_IO tomorrow_RT 's_$ impersonal_JJ intercontinental_JJ war_NN2 ,_, the_AT questions_NN2 seem_VV0 far-fetched_JJ *(3844 parses)*

Thus correct syntactic interpretations can be produced for the function of punctuation marks by the integration of relatively simple principles and rule schemata into existant grammars. Additionally, it seems possible to represent these principles in a relatively small number of concrete rules in the actual grammar, which is encouraging from the point of view of grammar size and complexity.

Having shown that the principles developed for the representation of the syntactic function of punctuation marks are appropriate, it is important to investigate the semantic contibution that punctuation marks can make to language analysis, and this contribution is explored in the next chapter.


```
(((She_PPHS1
   ((remained_VVD (squatting_VVG ((on_II (her_APP$ (heels_NN2))))))
    ((all_DB) (the_AT (time_NNT1 (we_PPIS2 (were_VBDR there_RL))))))
  ;_;
  (((like_ICS (the_AT (man_NN1)))) ,_,
   (she_PPHS1 (was_VBDZ ((((entirely_RR)) (naked_JJ)))))))
```

```
((((The_AT
    ((book_NN1 (concerned_VVN))
     ((with_IW
       (((the_AT (Negro_NN1)) 's_$)
        ((role_NN1) ((in_II (an_AT1 (((urban_JJ)) (society_NNJ1))))))))))
   ((is_VBZ ((rare_JJ))) ((indeed_RGA))))
  ;_;
  (((recently_RR))
   (((only_RR))
    ((((((Keith_NP1 (Wheeler_NP1))) 's_$) (novel_NN1)) ,_,
      ((((Peaceable_JJ)) (Lane_NNL1))) ,_,)
     (has_VHZ ((openly_RR)) (faced_VVD (the_AT (problem_NN1)))))))))
```

Figure 7.5: Appropriate parses of the semi-colon conjoined sentences

```
((((in_II (the_AT (((physical_JJ)) (sciences_NN2))))) ,_,
  ((these_DD2 (achievements_NN2))
   (concern_VV0
    ((electricity_NN1) ,_,
     ((chemistry_NN1) ,_, (and_CC (((atomic_JJ)) (physics_NN1)))))))))
```

```
(((((at_II (((((Jenks_NP1)) Street_NNL1)))) ,_,
    (((Simms_NP1)) (said_VVD))))
  ,_,
  ((the_AT (car_NN1))
   (((skidded_VVD) (completely_RR around_RL)) ,_,
    ((((just_RR)) (missed_VVN ((((two_MC)) (parked_VVN (cars_NN2)))))) ,_,
     (and_CC (sped_VVD (((East_ND1) ((in_II ((Jenks_NP1)))))))))))))
```

```
((((If_CS ((((one_MC1)) (dancer_NN1)) (slaps_VVZ another_DD1)))) ,_,
  ((the_AT (victim_NN1))
   (may_VM
    ((do_VD0 (a_AT1 (pirouette_NN1))) ,_,
     (((sit_VV0) ((down_RP))) ,_,
      (or_CC
       (offer_VV0 (his_APP$ (assailant_NN1))
        (a_AT1 ((fork_NN1 (and_CC spoon_NN1)))))))))))
```

```
((((to_II
    (the_AT
     ((men_NN2)
      ((in_II
        ((the_AT
          (instrument_NN1 (jamming_VVN (bomber_NN1 (cockpit_NN2))))) ,_,
         (((((submarine_JJ)) (compartment_NN2)))
          (and_CC
           (the_AT
            (((antiseptic_JJ)) ,_,
             (((windowless_JJ))
              (room_NN2 that_CST
               (will_VM
                (be_VB0
                 (the_AT
                  (foxhole_NN2
                   ((of_IO
                     ((tomorrow_RT 's_$)
                      (((impersonal_JJ))
                       (((intercontinental_JJ))
                        (war_NN2)))))))))))))))))))))) ,_,
  ((the_AT (questions_NN2)) (seem_VV0 ((far-fetched_JJ))))))
```

Figure 7.6: Appropriate parses for the sentences involving comma conjunction

eight The Role of Punctuation in Semantics

I walked back to the car. Every parked car along the street had a licence plate that said 'Missouri — the Show Me State'. I wondered idly if this could be short for 'Show Me the Way to Any Other State'. In any case, I crossed the Mississippi — still muddy, still strangely unimpressive — on a long, high bridge and turned my back on Missouri without regret. On the other side a sign said BUCKLE UP. ITS THE LAW IN ILLINOIS. *Just beyond it another said* AND WE STILL CAN'T PUNCTUATE. (Bryson, 1990)

One of the major problems involved in investigating the semantic and pragmatic facets of punctuation is deciding on the level of detail that should be adopted. Whilst almost all studies of punctuation have acknowledged that in addition to a syntactic function, punctuation also has a semantic role to play, there is a remarkable variety in the extent of those approaches. Nunberg, (1990), for example, acknowledges the presence of such a role, but excludes it from his principles on the grounds of being too context-dependent.

Dale, (1991), on the other hand, discusses the semantic role of punctuation in greater detail, and takes the view that one of the main semantic functions of punctuation marks is to act as signals of discourse structure. He suggests three possible roles for punctuation in discourse structure: to indicate the degree of rhetorical balance, the degree of aggregation, and to signal particular rhetorical relations (c.f. 4.1). However, he does acknowledge the point of context-dependence and underspecification made by Nunberg and others, and argues that very often the problem (in the field of language generation anyway) lies not in the precise function of the punctuation mark, but rather in the strength of punctuation appropriate. This concept can be extended to the field of language analysis to suggest that the problem lies in determining the specificity of meaning of a particular punctuation mark — e.g. does the comma in sentence (8.1)

indicate a null semantic link (i.e. just syntactic), indicate a simple sequential relation, or indicate a relation with a subtle distinction between the related elements.

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

Motivated by this analysis of Dale's, Briscoe (1996) implements a system that inserts discourse relations (at a very basic level) into a semantic analysis of the sentence. Based on the work of Lee (1995), who argues for this generalised approach as a solution to the problem of underdetermination, this system does little more than distinguish between coordinating and subordinating discourse relations and inserts one of these vastly underspecified relations into an analysis, as in (8.2). However, this still indicates that it is possible to use specific punctuation marks in particular sentential situations to determine discourse relations (at some level of specificity), and presumably a more complex system than this one might be able to further specify these relations, for example, by reference to context.

(8.2) a. The rumour — the Prince had been unfaithful — appeared in a newspaper.

b. **the**(x), **rumour**(x), **appear**(e,x), **a**(y), **newspaper**(y), **in**(e,y), **SubDR**(x,e'), **the**(z), **prince**(z), **be**(e',**unfaithful**(z))

In a more detailed and further-reaching investigation, Say and Akman (1996) discuss possible treatments of various aspects of the semantics of punctuation, and implement these treatments in Discourse Representation Theory (DRT) (Kamp and Reyle, 1993) using discourse representation structures (DRS's) and Asher's (1993) segmented DRS's (SDRS's) to present analyses. In addition to situating their punctuational analyses in the DRT paradigm, they further build on Dale's tripartite semantic punctuation-function hypothesis by illustrating not only instances where punctuation marks seem to encode specific rhetorical relations, but also examples where they indicate rhetorical balance and aggregation.

However, since Say and Akman are analysing the *information content* of the punctuation, they blend the syntactic and semantic roles of the punctuation marks. In the case of the sentences shown in (8.3 – 8.5), for example, and their respective DRS's (figures 8.1 and 8.2), the semantic phenomena illustrated seem to have nothing directly to do with the punctuation present in the sentences, but are simply based on the phrasal and clausal analyses of the sentences at syntactic level, which have, of course, in turn been influenced by the punctuation present. In relation to figures 8.1 and 8.2 it is worth stressing that in order to keep them simple, all the DRS's and SDRS's in this chapter omit complex tense information.

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

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

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

b. Tom has two cats, which once belonged to Fred, and Sam has one. (McCawley, 1981)

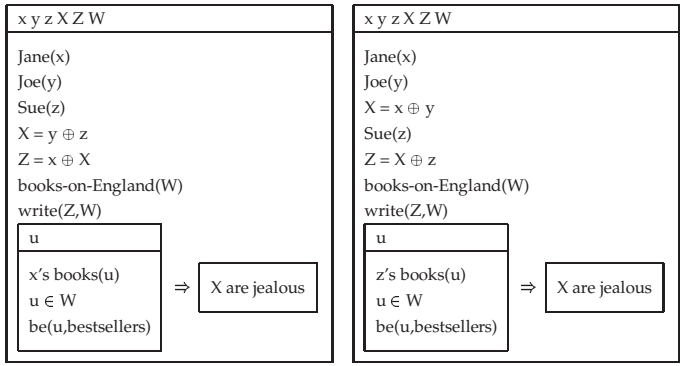


Figure 8.1: DRS's for (8.3), (Say and Akman, 1996)

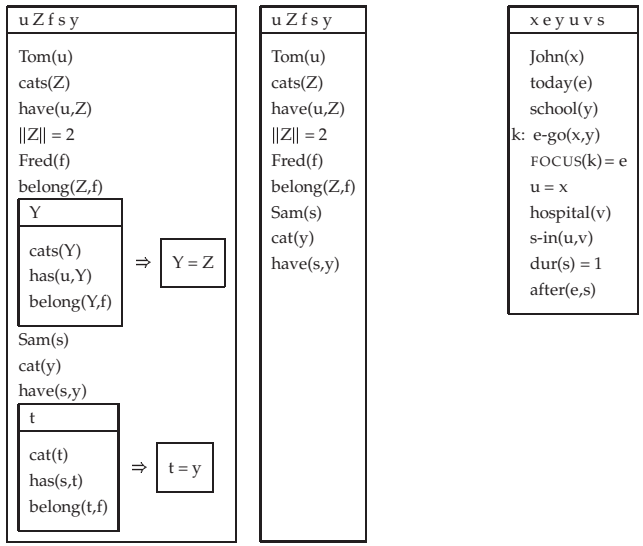


Figure 8.2: DRS's for (8.4) and (8.5), (Say and Akman, 1996)

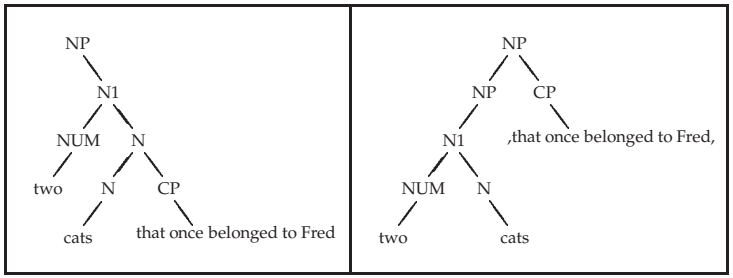


Figure 8.3: Possible parses for the object noun phrase in sentences (8.4)

(8.5) Today, John went to school. He has been hospitalised for a year. (Dawkins, 1995)

Thus in the sentences reproduced in (8.3) the position of the comma certainly determines the referents of the anaphors in the second sentence, but this is a process that happens at (or is at least is determined from information occurring at) the syntactic level i.e. the number and gender attributes of the noun phrases involved.

Similarly, the comma in sentence (8.4a) serves only to separate the two main constituent clauses of the sentence. This comma is actually optional, but fulfils a useful disambiguating role here due to the syntactic and semantic simailarity of *Fred* and *Sam* — without this comma it would be relatively easy to garden path this sentence. In a similar sentence that lacks this ambiguity (8.6) the comma is not necessary (although it would still be licensed). Conversely, the absence of a comma between the head object noun phrase *two cats* and the relative clause *that once belonged to Fred* tells us that the relative clause is restrictive on the head (Quirk et al., 1972). This is really a syntactic phenomenon rather than a semantic one, although it does depend on the particular grammar used as to whether the restrictive/non-restrictive distinction can be picked out at the parsing stage. It is, of course, the delimiting commas in (8.4b) that mark the relative clause as non-restrictive on the head.

(8.6) Tom has two cats that once belonged to Fred and they are both female.

Depending on the grammar used, the parse trees of the complete object noun phrases in (8.4) might be as shown in figure 8.3, and the difference between these parse structures should be enough to provide the distinction between the ultimate discourse structures. The DRS interpretations for the sentences in (8.4), shown in figure 8.2, are unsatisfactory in that they do not seem to capture the true distinction in the anaphoric references. Part of the reason for this is almost certainly the adoption of the simplified handling mechanism for plurals from (Asher, 1993).

Asher acknowledges suggestions that there should be a uniform type of discourse referent for plural entities and individuals, as discussed in (Frey and Kamp, 1987) and (Kamp and Reyle, 1993), but it is his simplistic treatment of plural discourse referents that provides problems for the interpretation of the sentences in (8.4). A related problem in these sentences is the use of

numerals such as *two* as anaphors. Thus we can provide more satisfactory SDRS's for (8.4) by using these anaphors as two-place predicates that isolate the requisite number of entities from the set of all such possible entities (c.f. figure 8.4). These SDRS's are shown in figure 8.5. Note that whereas the DRS of Say and Akman (1996) did not indicate the degree of rhetorical balance between the head object noun phrase and the relative clause in (8.4b), the delimiting punctuation in this sentence seems to require the insertion of a rhetorical relation into the SDRS, as in figure 8.5.

In sentence (8.5), similar to the other sentences explored so far, the phenomenon of focus on *today* is a perfectly valid one. However, this does not seem to be determined by the comma separating *today* from the rest of the sentence, or by the intonational boundary it coincides with, as claimed in (Say and Akman, 1996). Once again, this comma seems to be more important syntactically than semantically. A comma in this position classically serves to separate the fronting adverbial phrase from the rest of the sentence to avoid any phrase-boundary errors. In addition, it is possible to extract the focus information referred to in (Say and Akman, 1996) from the fronting position of *today*. There seems to be little difference in meaning between (8.7) and (8.8). Both focus *Today* in a way that (8.9) does not. In order to get the same focal emphasis in (8.9), the subsequent sentence would have to change from *He has been…* to *He had been….*

(8.7) Today, John went to school.

(8.8) Today John went to school.

(8.9) John went to school today.

However, this case is somehow less clear cut than the other two have been. The presence of the comma does seem to emphasise or confirm the focal role of *today*, especially since it is only a single-word phrase, although a comma in this position does not always fulfil this function. This illustrates a further question that emerges from the work of Say and Akman (and that has also emerged implicitly from other, less detailed work), namely which of the many punctuation marks encode some semantic/rhetorical/coherence relations, and when particular punctuation marks do encode such a relation, what is the identity of the relation.

Earlier work has tended to skirt this problem, or has declared it insoluble, by saying that the semantic functions of punctuation are too dependent on context and pragmatics for a rigorous treatment (Nunberg, 1990) or by underspecifying the relations encoded by punctuation marks

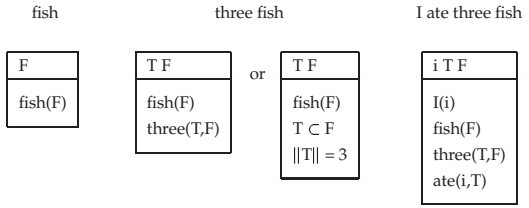


Figure 8.4: Illustrations of the function of numerical anaphora predicates

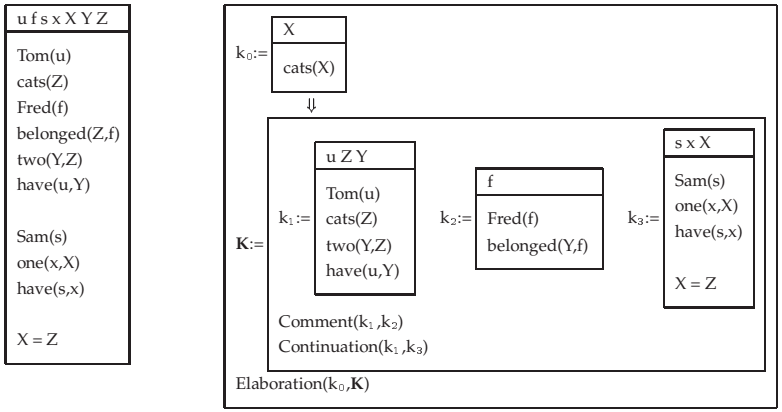


Figure 8.5: Proposed SDRS's for (8.4)

so far that they become almost useless (Briscoe, 1996). Say and Akman try to extract quite detailed relational information from the punctuation but in practice it is unlikely that information of such quality can be reliably achieved in any automatic or semi-automatic fashion, in all but a few cases.

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

In their analysis of sentence (8.10) (shown in figure 8.6) Say and Akman use the Parenthetical relation to alter the discourse significance (rhetorical balance) of the delimited phrase. This relation, an invention of theirs, is in effect very similar to the rhetorical relation Comment, with the obvious exception that Comment is a binary relation whereas Parenthetical is unary. However, the insertion of the DRS for this subordinate phrase into the SDRS that *elaborates* the topic of *John* ensures that it is still recognised as a modifying phrase, but one of lesser significance. To capture the true spirit of Asher's (1993) Topic-Based Updating, the SDRS for (8.10) might be better restated as in figure 8.7. Note, however, that it would be wrong to assume that it is the sentence boundary (full-stop and word-initial capitalisation) that is specifying for the relation Continuation, as the following sentence could equally well concern some entirely different topic. Note also that once again, for simplicity, figure 8.7 omits all temporal information.

An additional effect that the delimiting punctuation has here is to restrict the possible referents for anaphors. Thus in figure 8.7 the singular male referent of '**t**' could have been either John or his brother. However, the Comment relation, provided by the punctuation in the sentence, seems to dis-prefer the brother as the referent. If the sentence were as in (8.11), however, the referent of the pronoun in the second sentence is clearly the sister. This shows us that the Comment relation does not hide its argument from the anaphoric resolution process

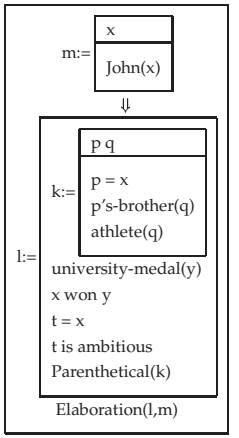


Figure 8.6: SDRS for (8.10), (Say and Akman, 1996)

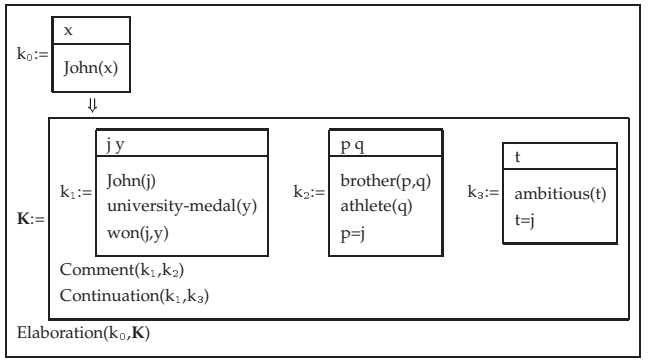


Figure 8.7: Proposed SDRS for (8.10)

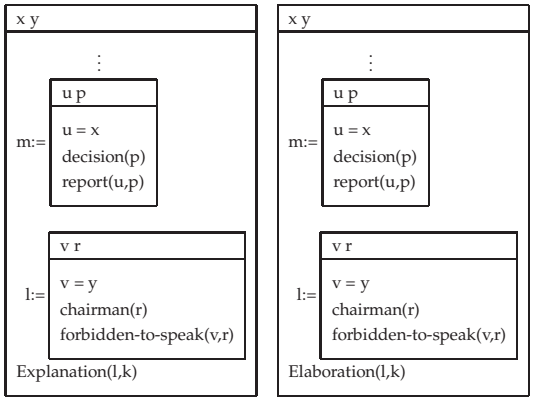


Figure 8.8: SDRS's for (8.12), (Say and Akman,1996)

completely.

(8.11) John — his sister also an athlete — won the medal. She had taught him well.

(8.12) 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.

A more contentious example given by (Say and Akman, 1996) is one borrowed from (Nunberg, 1990). They give the SDRS's of sentences (8.12) as in figure 8.8. Nunberg himself describes the second clause in each of these sentences as being interpreted to be an elaboration of what is decided and an explanation of why the decision is reported as it is, respectively. Therefore Say and Akman appear to have the coherence relations the wrong way round in the two SDRS's. Hobbs (1985), in his description of coherence relations, describes the Explanation relation as one that allows us to infer that

> the state or event asserted by a discourse fragment $\mathcal{S}_1$ causes or could cause the state or event asserted by a preceding discourse fragment $\mathcal{S}_0$.

Thus the correct discourse relation for the second sentence in (8.12) is certainly Explanation. The description of the relation Elaboration in (Hobbs, 1985) is somewhat less satisfying for our purposes, however. He defines it as allowing us to

> infer the same proposition $\mathcal{P}$ from the assertions of discourse fragments $\mathcal{S}_0$ and $\mathcal{S}_1$, the latter frequently, but not necessarily, supplying additional, crucial information.

In the case of the first sentence in (8.12) the second clause is indeed performing an elaborating or defining role, but not on the entirety of the first clause. The colon-expansion is only describing

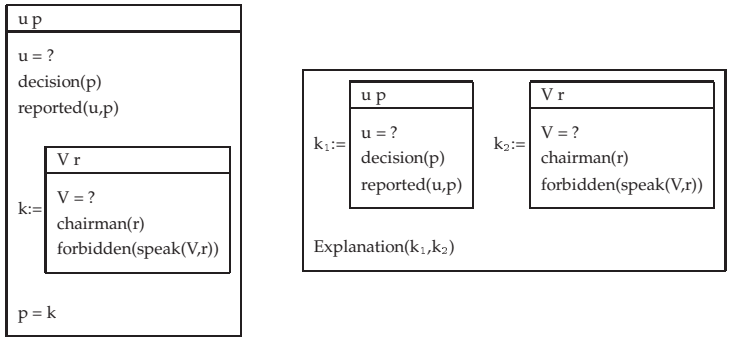


Figure 8.9: Proposed SDRS's for (8.12)

the *decision*, so the Elaboration relation hardly seems appropriate since it would refer to the elaboration of the entire discourse fragment preceding the colon, and even if it could be made to refer to just the *decision*, it is hardly an *elaboration*, but rather a *definition*. Therefore some other means must be found of representing this sentence in the discourse structure. More suitable representations might be as shown in figure 8.9. Even this analysis may not be dependent on the punctuation present, however. The presence of a semi-colon is sufficient to signal that the elaboration or description associated with the colon does not apply in this instance, but the semi-colon does not uniquely specify or entail the Explanation relation; this relation is specified more by the context than the punctuation mark.

From these, and other examples, it becomes clear that there is a certain amount of semantic information held by punctuation marks, but there is wide-spread disagreement as to the extent and usefulness of this information. Approaches that discount punctuation's semantic content as too contextually or pragmatically specified are, quite clearly, unable to make any use whatsoever of information that is present, whereas those approaches that try to extract information at too specific a level run the risk, as we have seen, of extracting the wrong information, or extracting information that is contentious.

It seems necessary to classify and describe the punctuation marks that can reliably be used to specify particular semantic functions, and what those functions might be. As so often in the field of semantics, use will almost certainly have to be made of *sets* of possible interpretations or generalisations, which can be selected for or specified at a later stage using context or pragmatic knowledge (c.f. the approaches adopted in the interpretation of nominal compounds (Jones, 1995b; Jones, 1995d; Johnston et al., 1995; Johnston and Busa, 1996)). Additionally, such semantic paradigms as defeasible default interpretations (Lascarides et al., 1996), will almost certainly have to be adopted.

8.1 Classification of Semantic Functions

There are, evidently, several instantly obvious categories into which the semantic function of punctuation marks can be divided. These include the tripartite classification in (Dale, 1991), but also include other categories. The complete set of categories will be roughly as follows:

- **Null function** — obviously there are certain punctuation marks that serve a purpose that is purely syntactic. These therefore have no semantic role whatsoever. Also included in this category will be those marks whose possible semantic interpretations are so general as to be entirely useless.
- **Lexical entities** — certain non-alphanumeric marks have specific lexical meaning, which might well be dependent on the genre and context of the text in which they appear. These marks should therefore be treated similarly to other lexicalised objects e.g. words.
- **Discursive functions** — as discussed in (Dale, 1991)
  - **Balance** — marks which affect and determine the rhetorical balance of a discourse.
  - **Aggregation** — marks which affect and determine the aggregation of the discourse fragments.
  - **Relational** — certain marks will specify particular, or a set of, discourse relations. These could be *unary* relations (for example, as specified by stress markers) or *binary* and higher-arity relations (the more conventional uses we have already encountered in this chapter).

It is highly likely that certain marks of punctuation will fall into more than one category, for example affecting the degree of rhetorical balance as well as specifying a set of rhetorical relations. The following is not intended to be an exhaustive list of all possible functions of punctuation, but instead to be a guide for likely semantic functions of punctuation marks used in usual circumstances.

Null Function

Many punctuation marks have no semantic relevance whatsoever. Their function is either entirely at a syntactic level, or some other even earlier level of processing (e.g. morphology). The first set of punctuation to fall into this category is *sub-lexical* punctuation (8.13). This punctuation is used to drive morphological and lexical look-up processes, but does not carry forward any information to further analysis stages. The mark that could possibly be argued to be an exception to this is the word-final possessive apostrophe (8.14). Here, the only thing that signals the possessive attribute is the punctuation mark itself, so it could be argued that it carries the same semantic content as a normal, lexical, possessive suffix (as in (8.13)). However, the possessive nature of the noun must be made apparent to the syntactic processing stage in any linguistic analysis, so any morphological system must convert the apostrophe into a possessive suffix marker (8.15), and so the punctuation mark as an entity is lost for the purposes of further analysis.

(8.13) down-sizing children's isn't I.B.M.

(8.14) teachers' pay claims

(8.15) ≪teacher≫ ≪+plural≫ ≪+poss≫

Many of the marks we associate more conventionally with the term *punctuation* — the *inter-lexical* marks — also have little or no semantic function. The use of the comma that Nunberg (1990) refers to as *disambiguating*, and that has been reclassified in section 6.2 as an adjunct-final marker, clearly serves no semantic function, since its only role and purpose is to disambiguate sentences from any garden path readings (8.16) and to indicate the precise clausal structure of the sentence (8.17).

(8.16) Those who can, contribute to the fund.

(8.17) The House of Commons having passed the third reading by a large majority after an animated debate, the bill was sent to the Lords. (Jarvie, 1992)

Another, similar, instance of punctuation without semantic function is in the case of pre-conjunction punctuation marks, where the function of the punctuation marks serves only at the syntactic level to indicate and disambiguate the clausal structure of the sentence (8.18–8.20). It is worth noting however, that only singleton marks can be treated in this way — pre-conjunct marks that stem from some other source, such as final delimiting marks, cannot be treated as semantically void (at least by this principle). In addition, usually only the marks capable of serving as conjuncts (i.e. comma and semi-colon) can function in this manner. However, in certain writing styles other singleton 'point' marks that appear in this position can be treated in a similar manner, which greatly facilitates the analysis of sentences such as (8.21) and (8.22).

(8.18) She was sick, and tired of studying punctuation. (Jarvie, 1992)

(8.19) She was a highly intelligent woman, but she seems to have have acted stupidly on this occasion.

(8.20) Among the speakers were John, a linguist; Mary, a lawyer; and Ed. (Nunberg, 1990)

(8.21) She had cried, she had implored, she had been miserable at this refusal, and finally he had relented — and now how happy she was, how expectant. (Say and Akman, 1995)

(8.22) Now, I tell you the entire story — but first you have another cup of coffee. (Say and Akman, 1995)

In their analysis of (8.21), Say and Akman (1995) suggest that there is a different meaning of the dash to the more conventional parenthetical interpretation of the dash interpolation, namely one that suggest that the fragment following the dash is a result of the discourse fragments coming before it, i.e. there is a Cause relation operating. This does indeed seem to be the case, but it is unlikely that the dash specifies this relation; rather it is the content of the particular discourse fragments that suggests this, and the dash itself does not seem to have any

semantic contribution to the discourse, only a syntactic one in separating the sentence into two multi-clausal constituent fragments. This is reinforced by the examination of (8.22). In (Say and Akman, 1995), it is suggested that the use of the dash in this sentence is a return to the conventional parenthetical dash interpolation. This does not seem to be the case, however, as the second clause seems to function at a higher sentential level than a mere comment — the latter as illustrated in sentence (8.10). Neither, however, should the sentence be treated in the way that Say and Akman treat (8.21), since there does not seem to be a Cause relation operating, or indeed any other particular discourse relation. Therefore it seems most suitable to treat these dashes as syntactically functional, but semantically void.

The final instance of punctuation marks that are without a semantic role are all those marks that complete delimited phrases. As we will see later on in this chapter, the initial mark of a delimited phrase indicates not only the rhetorical balance (i.e. that the delimited phrase is subordinate to the rest of the sentence) but can also specify particular discourse relations. However, the sole function of the closing punctuation mark, if there is one present, is to mark the completion of the subordinate phrase to the syntactic analysis. There is no need, at a discourse level, for the mark to indicate a 'return to the outer discourse fragment' since the extent and separation of the subordinate fragment will have been marked during parsing.

As an aside, the highest level delimiting punctuation functions that are likely to be encountered in normal text, namely the sentential ones, are also included in this principle. Thus the full stop (the complementary punctuation mark to the sentence-initial capitalisation) can be treated as semantically void since it is the final delimiter. Obviously, in cases where the full-stop is absorbed by a sentence-final stress marker, the semantic content of that mark will over-ride the conventional semantically empty nature of that position.

It might reasonably be expected that the sentence initial capitalisation might carry semantic information of an equal weight to other initial delimiting marks. However, it is difficult to determine exactly what the nature of the contribution of the semantics of this phenomenon might be. Often, a new sentence indicates a Continuation relation, but it can equally well signal almost any other relation, depending on the context, including Elaboration, Explanation and Parallel, or can signal no relation at all. Therefore it seems most suitable to declare this punctuation function semantically empty and to try to extract semantic relations that tie together discourse structures for adjacent sentences from the context and content of those sentences rather than from the marks that delimit them.

Lexical Function

Many non-alphanumeric marks simply represent normal lexical items, and therefore will have similar syntactic and semantic content to other, more conventional lexical items, such as words. The most frequent set of marks that fall into this category are those lower frequency, genre-specific marks described in section 5.3, whose precise identity and use tends to change depending on the text source being used, e.g. (8.23).

(8.23) **# % & ∗ $;-) + =**

These should just be treated as other lexical items are, in terms of syntactic and semantic analysis. The only punctuation mark that is not in this set, but falls into the category of a fixed lexical item, is the inter-lexical punctuation mark for ellipsis (…). The most simple treatment for this mark is to lexicalise it as *etcetera*, *and so on*, or other similar terms; alternatively in certain circumstances it may be more suitable to replace the ellipsis with some underspecified item signifying that there is some lexical information not present in the text, or that there is a pause or interruption in the flow of text. Note that it is important to distinguish this standard punctuational usage of the ellipsis from usages in quoted speech (where it signifies that some more material was originally spoken but is not present in the current representation, or again a pause or interruption) and in square brackets ([…]) where the entire punctuational aggregate (including square brackets) is an editing-related symbol indicating deleted material. An important thing to watch for is the possible interaction of the ellipsis with other, graphically similar marks, in terms of absorption which is, of course, not something that lexical items usually perform.

Two other inter-lexical punctuation marks should also be included here, although their use is not so much as lexical items as it is as anaphors (cataphors actually). These are the commas or semi-colons that separate lists, as in (8.24) and (8.25). In the former, the commas are resolved to the conjunction *and*, and in the latter they are resolved to the disjunction *or*. In addition, these marks can occur exophorically, i.e. on their own without any coordination to refer to, as in (8.26),(8.27) and (8.28). In these circumstances the default lexicalisation and function for these marks should be as a conjunction. In (8.29) the conjunction *but* seems more appropriate than *and*, and so the default could be over-ridden by the context (of the second clause contrasting with the first one).

(8.24) We bought books, tapes, CD's and T-shirts.

(8.25) Are we going to fly to Rome, drive to London or walk to Edinburgh?

(8.26) I will come on foot, you should drive.

(8.27) It was a tall, ugly, post-war municipal building. (Jarvie, 1992)

(8.28) He was a distinguished academic geographer; he had wide-ranging research skills; he published books in all sorts of subject areas; he was also a devoted husband and father. (Jarvie, 1992)

(8.29) We liked John; we disliked his politics. (Jarvie, 1992)

Depending on genre, other punctuation marks can also serve in this role. For example in letter-writing it is very common to see 'stream of consciousness' style being employed, where subsequent phrases are written down as they come to mind and are linked with dashes. Here, the dash is non-delimiting, and is fulfilling a low-level conjoining role: *and* is maybe a little too strong as a lexicalisation, but some low level connection is implied by the use of these marks.

Discursive Functions — Rhetorical Balance

Rhetorical balance is a term that refers to the relative importance of juxtaposed elements in the discourse, for example, the distinction between *nuclear* and *satellite* material: the former being the key elements of the discourse, and the latter the less important elements that could be deleted without the sense of the text being affected. This is not to say that the content of the satellite material is of less value, indeed, often the satellite phrases have a richer content than the head phrase, in the case of elaborations, for example.

It is instantly apparent that many punctuation marks fulfil the role of distinguishing nuclear material from satellite. Any delimiting and adjunctive structure (comma-delimited adjuncts, colon-expansion, dash-interpolation, parenthesis etc.) performs this function, declaring the delimited item to be of lower importance than the surrounding discourse.

However, in a sense this distinction has already been made at the syntactic processing level, as the delimited items will have already been analysed as non-restrictive modifying clauses, and it is this distinction that determines the rhetorical balance between the head phrase and the modifying clause. The distinction between semantics and syntax becomes very confused at this point, however, since very often the only way of signalling a non-restrictive modifier is to surround it with punctuation marks of some description. Thus the nature of the function of these punctuation marks depends entirely on the design of the grammar, and whether the punctuation marks are used to produce an analysis of the modifying phrases that declares them non-restrictive, or whether the phrases are treated the same as other, restrictive modifiers, in which case the non-restrictive distinction must be made at the semantic level.

For neatness, and since I would like to think that, certainly in some cases, the parse trees resulting from syntactic analysis will reflect the non-modifying nature of punctuation-delimited phrases (as in figure 8.3), I believe that the signalling by punctuation of rhetorical balance for discourse analysis purposes will happen via prior syntactic analysis, and hence there is no specifically semantic role for punctuation to play in this category.

There is, however, one phenomenon that affects rhetorical balance that cannot be decribed satisfactorily at a syntactic level. This relates to the relative importance or unimportance of punctuationally delimited items. Sentences (8.30–8.33) all make the same statement, but the relative importance of the delimited clause becomes smaller in each successive example. The additional information, that *Fred* is the President's brother, is relatively immediate to the discourse in (8.30), but becomes less so through (8.31) and (8.32) until it is hardly relevant at all in (8.33). The precise level of the immediacy of the subordinate clause is not really important in the examples shown, but in a situation where several of these types of phenomenon occur in the same discourse, it will be important to be able to assess which are of greater significance to the discourse.

(8.30) Fred, the brother of the President, was sent to jail.

(8.31) Fred — the brother of the President — was sent to jail.

(8.32) Fred (the brother of the President) was sent to jail.

(8.33) Fred[1] was sent to jail.

Thus the semantic category of rhetorical balance, as signalled by punctuation, is not entirely empty. Delimiting punctuation has the power to signal the degree of importance of the delimited discourse fragment to the surrounding discourse, via a crude power hierarchy (8.34).

(8.34) , ≫ — ≫ ( ) ≫ footnotes etc.

Note that it is likely to be difficult to establish any absolute measure of relevance or importance from these particular punctuation marks; their use will only be able to determine whether one type of punctuation-delimited discourse is of greater relevance to the overall discourse than another, differently delimited discourse.

Discursive Functions — Aggregation

At first sight, the situation with the signalling by punctuation of aggregation (the notion of the relative closeness and distance of juxtaposed material in the discourse) would appear to be very similar to that of rhetorical balance, except of course for the disinction that whereas rhetorical balance seems to be signalled only by delimiting punctuation, aggregation is signalled only by conjunctive punctuation (or separating punctuation using the terminology of (Nunberg, 1990)).

The fact that sentence (8.21) has two main constituents, the first of which is subdivided into four parts and the second into two, is due to the recognition of the greater *power* of the dash as an element of punctuation over the comma. If we imagine a further addition of complexity to the sentence, as in (8.35), it is the greater power of the semi-colon over the dash (and therefore over the comma) that determines the new aggregation of the sentence into three portions, the first of which has a similar sub-structure (since it is the same sentence) to that of (8.21), discussed earlier.

(8.35) She had cried, she had implored, she had been miserable at this refusal, and finally he had relented — and now how happy she was, how expectant; she had had terrible difficulties in the past trying to get her journal articles accepted by this reviewer; her standing in the academic world was certain to improve with this newest development.

This aggregational analysis, however, is one that should really occur at the parsing level — the parse tree for (8.35) would certainly reflect the top-level tripartite structure and the other deeper structural phenomena, and so there does not really seem to be a role for punctuational signalling of aggregation as a semantic phenomenon. Any indication of aggregation will be accessible from the syntactic analysis of the sentence. Therefore it seems that once again there is no semantic role for punctuation to play in this category.

It is interesting that this holds true for higher level punctuational phenomena also. Super-lexical punctuation such as structure and spacing performs similar aggregating roles, at a

[1] The brother of the President

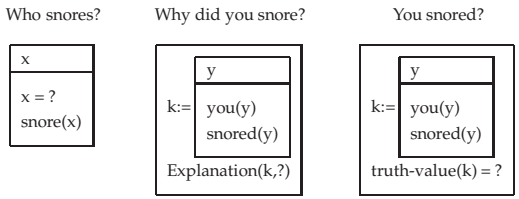


Figure 8.10: Possible discursive functions of the question mark

higher level, to the marks of inter-lexical punctuation discussed already. These could also be used at a syntactic processing level if the phenomena are described by some suitable mark-up language. The hierarchy of *power* of punctuation phenomena might then resemble that given in (8.36).

(8.36) chapter ≫ section ≫ grouping ≫ paragraph ≫ sentence ≫ ; ≫ — ≫ ,

Discursive Functions — Discourse Relations

This is the semantic function of punctuation that is most commonly used and referred to, but as we have seen the variety of approaches differs greatly. It seems certain that some punctuation marks are able to signal specific discourse relations, or at least a small set of possible relations, whereas others equally certainly underspecify discourse relations to the extent that trying to extract any semantic information from them is a futile exercise.

The simplest, and most reliable marks as regards signalling of discourse relations are the two most common stress markers (the third, ellipsis, as we have seen is best treated by lexicalisation or replacement with some underspecified structure). The question-mark simply specifies that some portion of the immediately preceding discourse is a question, which admits several different treatments in discourse structure, as in figure 8.10. Of course, some sentences that terminate with a question mark can be recognised as questions without the presence of the stress marker, due to the presence of some *wh*-word or phrase. Thus the importance of the question mark for discourse analysis is reduced, but it is still necessary in many cases, as in the third example in figure 8.10. The exclamation mark allows the preceding portion of the discourse to be emphasized in terms of importance to the discourse, strength of individual belief, or truth value.

Apart from the stress markers, the punctuation mark that is most usually associated with a discourse relation role is the colon. It is defined as:

> The colon is the preferred punctuation mark for introducing something which explains or illustrates or rephrases a previous statement: a quotation, a list, an idea, a definition, an example, etc. (Jarvie, 1992)

There are two problems with a colon signalling discourse relations, however. The first is the precise identity of such a relation — it could reasonably be one, or several, or even none of

Elaboration, Explanation, Parallel and Exemplification. The second problem is to determine the referents of the relation once it has been chosen. Obviously, the discourse fragment contained in the colon expansion will be one of the referents, but the other referent is not necessarily the entire discourse fragment before the colon. In some examples, as in (8.12) only the object noun phrase of the sentence before the colon is referred to by the colon-expansion, and this will be important to the resolution of the final discourse structure. In figure 8.9, for example, since the colon-expansion elaborates on the noun phrase rather than the entire discourse, the best solution was not to use a discourse relation at all, but to insert an additional equality to describe the content of the decision. If available, some additional discourse relation such as Content could have been used.

Therefore, the most appropriate course of action seems to be to associate the colon (via defeasible unification (Lascarides et al., 1996) in this description) with its most likely discourse relation, Elaboration (8.37). Alternatively, if it is preferable to produce a set of possible relations (as is the case with nominal compounds, for example (Jones, 1995b)), then such a set could be defeasibly specified e.g. (8.38).

(8.37) $\text{Relation}_{colon} = /\text{Elaboration}(\mathcal{S}_0,\mathcal{S}_1)$

(8.38) $\text{Relation}_{colon} = /\{\text{Elaboration}(\mathcal{S}_0,\mathcal{S}_1), \text{Explanation}(\mathcal{S}_0,\mathcal{S}_1), \text{Exemplification}(\mathcal{S}_0,\mathcal{S}_1)\}$

Choosing whether to accept the default interpretations, or specifying alternatives, is a task which must be carried out at a context-specific or pragmatic level, and as such is outside the realms of the investigation of punctuation *per se*.

The other common instance of default relations specified by punctuation is when delimiting punctuation surrounds a non-restrictive modifying discourse fragment. As we saw in the treatment of sentence (8.10), as shown in figure 8.7, delimited phrases can always be related to surrounding discourse via, at the most basic level, a Comment relation. If the analysis is being carried out in the spirit of Asher's (1993) Topic-Based Updating, then this is the only relational information the punctuation needs to contribute specifically, since the other information that is carried, namely the notion of elaboration of the topic that precedes the delimited discourse fragment, is already represented by the Elaboration relation inherent with the topic-based updating paradigm. Hence the representation of the complex delimiter-containing sentence (8.39) might be as in figure 8.11.

(8.39) John — a neighbour of Fred, who hates him — was away.

The only further relational role triggered by specific marks of punctuation seems to be with quotation marks. Narrative quotation (8.40) will obviously trigger the Narrative relation, and scare quotes should be handled in some system-specific manner. One possible use of scare quotes is to relax semantic and pragmatic constraints on the lexical items they surround, so that unusual words can be accepted in this context. Alternatively they could signal that the lexical items they surround should receive some metaphorical interpretation.

(8.40) John replied that he "had not known at the time."

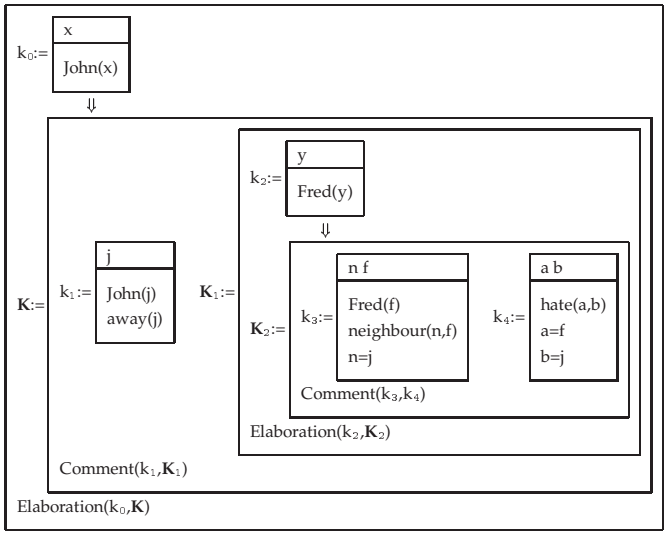


Figure 8.11: Proposed SDRS for (8.39)

8.2 Summary

The underspecified nature of semantic information contained by punctuation has not provided as many problems to this investigation as the question of whether the information used is actually strictly semantic or syntactic. Following the assumptions made here, we seem to be able to ignore a large part of the punctuation marks as regards semantic content.

A specific class of punctuation marks lend themselves to a lexicalised interpretation, with the obvious rich and varied semantic content this implies. This class includes the source-dependent variable set of non-standard punctuation marks (as in section 5.3), conjunctive punctuation marks (separating, under the scheme in (Nunberg, 1990)), and a small number of other marks.

Rhetorical balance does not seem to be influenced by punctuation at a semantic level, at least not directly. The information that punctuation does contribute is structuralised at the syntactic processing stage and the semantic information is extracted from the nature of this structure.

In a similar fashion, aggregation, as regards conjunctive punctuation and adjacent discourse fragments, is also an operation that is determined from punctuation at a syntactic level. However, the relative importance of delimited discourse fragments to the text and to one another can be determined via a power hierarchy of punctuation marks and phenomena.

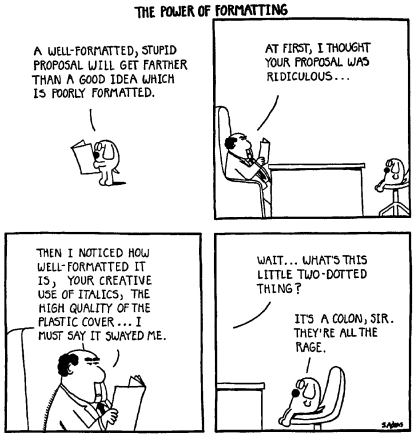
The most frequently considered semantic function of punctuation — marking for specific discourse relations — is only of use with a limited set of marks. The stress markers, the colon and quotation marks all are able to encode discourse relations, but the precise identity and

applicability of these relations must be determined with reference to context and pragmatics.

Delimiting punctuation, however, seems to be able to consistently encode a Comment relation which if represented in the topic-based updating paradigm is also included in a discourse entity that modifies the discourse topic via an Elaboration relation.

Additionally, and not necessarily directly related to semantics or discourse, but more related to anaphora resolution, the punctuation marks and their semantic effects do seem to be able to influence the process of choosing anaphoric referents. Possible referents in delimited discourse fragments are dis-preferred as referents of later anaphors, although they are not removed completely as possibilities. Only if no suitable candidate emerges from the discourse outside the delimitation, are the referents inside the delimitation brought into consideration.

nine A Taxonomy of Punctuation



(Adams, 1992)

Using the information presented in the preceding chapters, it is now possible to integrate all of the data on the identity, form, variety and function of punctuation marks into a taxonomy, in order to categorise the marks.

In actual fact, it seems that the taxonomy of punctuation functions can be expressed best via a hierarchy of the various categories. The first categorisation to be made involves the specification of what constitutes *punctuation* and how this is differentiated from the rest of orthographic material.

Orthographic Entities

The class of orthographic material we are likely to encounter can be subdivided into broadly four categories:

- **Lexical Entities** — these are classically regarded as the core constituents of orthography, carrying the bulk of the linguistic information. Lexical entities in orthographic text are most usually simple words or word-fragments (morphemes), and are usually composed solely of alphabetical characters. Their function and meaning can usually be obtained directly from the lexicon.
- **Numerical Entities** — these can occur throughout the text in conjunction with, or separate from, the normal lexical entities. Used to convey dates, amounts or other numerical data, these entities usually need to be analysed by a simple parsing process to determine their meaning (e.g. the number 42 is unlikely to be lexicalised, so it must be parsed to yield the interpretation "$4 \times 10 + 2$"). These entities are most usually composed entirely out of numerical characters.
- **Graphical Entities** — these are pictures that occur in the text and contribute to the overall information content. Usually, these are very difficult to process automatically. Note that this category does not include tables or other similar entities that consist solely of spatially arranged lexical and numerical items.
- **Punctuation** — having defined punctuation as that category that includes all orthographical information that is not alphanumeric or specifically graphical, this category naturally accounts for all other orthographic phenomena. These phenomena can range from printed characters (such as the decimal point or full stop) to the overall structure of the page or document, as determined by spacing, font size etc.

Punctuation

As discussed in section 1.1, and as exemplified in subsequent chapters, the category of punctuation is easily subdivided further, dependent on the position and nature of the punctuation phenomena in relation to other orthographic entities (lexical, numerical, etc.).

- **Sub-lexical phenomena** — these are the punctuation marks that classically appear within lexical and numerical entities, and therefore include such items as the decimal point, the hyphen and the apostrophe (9.1).

  (9.1) L.S.D. X-ray it's Joe's 12.325 1.4367×10^{-4}

  Their use tends to be well-understood and easily formally representable, thus making analysis and processing of these punctuation marks relatively simple. Problems are most likely to arise in those cases where a sub-lexical mark is orthographically similar to an inter-lexical punctuation mark and occurs at an extremity of a lexical or numerical item rather than within it. In these cases, it is possible for sub- and inter-lexical punctuation to

become confused and ambiguous, and in certain cases the marks interact so as to remove one or other of them (9.2).

(9.2) ≪Fujitsu bought I.C.L.≫ ≪full-stop≫ ⇒ Fujitsu bought I.C.L.

- **Inter-lexical phenomena** — these phenomena include those marks that are most conventionally regarded as constituting punctuation, but also include any other non-alphanumerical characters and character composities that occur between or around the lexical or numerical orthographic entities (9.3).

  (9.3) **. , ; : — ( ) “ ” ? ! @ = + # £ & \ ;-) +-**

  This category of punctuation also includes non-character phenomena such as sentence-initial word capitalisation. It is this category that contains the richest exploitable set of punctuation phenomena for the purposes of computational analysis and generation, and is therefore the one that has been focussed on most in the current work.

- **Super-lexical phenomena** — performing, in certain cases, similar functions to the inter-lexical phenomena, this category of punctuation includes all the phenomena that happen at a higher level than that of lexical and numerical items. This therefore includes all the other orthographic processes present in text representation: paragraphing, separation of text elements by means of whitespace, text-subdivision into sections and chapters, font- and typeface-alternation, underlining etc.

  Although a potentially rich source of information for natural language processing, it is currently difficult to exploit this category of structural punctuation due to the absence of a means of representing structural information (such as SGML or LaTeX) emerging and being used as a standard, and more particularly due to a lack of texts rendered into such a representation.

Classes of Inter-lexical Punctuation

As shown in chapter 5, and subsequently, inter-lexical punctuation can be divided into two rough categories, which differ in terms of function, interpretation and size.

- **Source-specific punctuation** — this is a set of punctuation marks, phenomena and mark-composites whose contents and interpretation varies dependent on the material being analysed.

  The basic constituency of this set can be described as containing all the punctuation phenomena not covered by the following category of *source-independent punctuation*. Therefore this category contains all the unusual punctuation symbols that either do not have a standard interpretation across all texts or are unlikely to occur in the majority of texts (9.4).

  (9.4) **$ # @ ;-) _ = +** ♡ ∞

As discussed in chapter 8, the most usual method of processing these punctuation marks is for them to be lexicalised (e.g. $ @), so that lexical look-up retrieves a complete structure describing the syntactic and semantic function of the particular punctuation mark or marks. However, this is not always the case: certain of these marks are likely to have relational functionality, i.e. they delimit some other lexical or numerical item to influence its interpretation (9.5) or they relate the adjacent lexical or numerical items in some particular manner (9.6).

(9.5) He was a ∗real∗ friend to me.

(9.6) 2 + 2 = 4

In these cases, the particular relational functionality of these phenomena must be encoded in the processing system alongside the functionality of the more conventional , source-independent punctuation marks. In some cases, such as (9.5), it might be possible to infer the function of a novel relational punctuation phenomenon without specific mechanisms to handle it being in place, but in general this is unlikely to be the case.

Processing of these source-dependent punctuation phenomena, then, relies on foreknowledge of the particular phenomena present in the text source to be analysed, and an encoding within the system of the interpretation to applied to these phenomena, which is likely to change with a different text-style.

- **Source-independent punctuation** — this is the set of punctuation marks and processes that can reasonably be expected to occur across all texts, and for which a standard interpretation can be developed that can be used across all texts to be processed. The set includes all the standard marks of point-punctuation, as defined in (Nunberg, 1990), as well as a particular variety of quotation and parenthesis operations, which will be source-dependent for typographical form, but which will have a standard, generalisable treatment. Additionally containing the process of sentence-initial capitalisation (as a partnering delimiter to the full-stop), the set will be as in (9.7), with the proviso that the actual orthographic manifestation of the quotation marks and parentheses could vary, as in (9.8).

  (9.7) **. , ; : ? ! — ( ) “ ”**

  (9.8) **“ ” ‘ ’ ' " () []** { }

  These marks all have syntactic and semantic roles to play in the analysis and generation of text, and so computational treatments can incorporate a standard set of interpretations for these marks. Classically, they are very rich syntactically, and can greatly assist in the parsing process (since they almost invariably mark phrasal boundaries of some sort). Semantically, the marks are less rich than the lexicalised source-specific ones, and tend to have relational interpretations.

Categories of Source-independent Punctuation

This informationally rich class of punctuation marks and phenomena can be subdivided into three categories, depending on the pattern of occurrence of the marks. The categories into which punctuation phenomena fall determine their broad syntactic and semantic functionalities.

- **Adjunctive punctuation** — approximating the class of marks described in (Nunberg, 1990) as *delimiting* punctuation.

Form

This category of punctuation marks and phenomena consists of paired punctuation entities occurring to either side of a set of non-punctuation orthographic items (e.g. lexical items) (9.9).

(9.9) **,** with a drink**,** **(**except for viewers in Northern Ireland**)**

(9.10) In the beginning, there was light.

(9.11) It was my closest friend, the President!

(9.12) Present were John, a teacher; Mary, a cook; and Fred.

According to the principles of point absorption set down in (Nunberg, 1990) and the principle controlling adjunctive punctuation set out in chapter 6 either of these marks can be missing due to co-occurrence with a more powerful punctuation entity. In (9.10), the initial comma surrounding the phrase *in the beginning* is absorbed by the phenomenon of sentence-initial capitalisation (C); conversely, the final delimiting comma is absorbed by the exclamation mark in (9.11). The more powerful entity that can absorb a delimiting mark does not necessarily need to be involved in an adjunctive pairing — it can also be a member of a different punctuational category. In example (9.12), for instance, the final delimiting commas are absorbed by non-delimiting semi-colons. The hierarchy for absorption power seems to be as in (9.13) for initial delimiter absorption, and (9.14) for final delimiter absorption. The reason why the stress markers ? and ! are not mentioned in (9.14) is that these can co-occur with less-powerful marks in a midsentence position (9.15). The only reason these marks seem to absorb delimiters in sentence-final positions is that the full-stop that marks the end of the sentence actually performs the obsorbtion, and then is itself absorbed via *graphic absorption* by the stress marker.

(9.13) C (" ; : ≫ — ≫ ,

(9.14) . ...) ; : ≫ — ≫ ,

(9.15) Is he here?, will they appear?

As we can see, dissimilar marks can form a delimiting pairing (as with sentence-initial capitalisation and sentence-final full-stopping). Another example of this is the colon-expansion. Here, the colon is the initial delimiting mark and the final delimiter is a

null mark that lies between the first and second categories in (9.14) in power. Thus the amount of text delimited by a colon-expansion extends either to the end of the sentence or to a semi-colon. Note that the situation here is complicated in terms of an absorption hierarchy since a semi-colon need not necessarily absorb the closure marker of the colon-expansion: it is also possible for the colon-expansion to scope over the semi-colon. The reason that the colon is treated as a delimiter and not in the other main category (separating/conjunctive) is for the scoping reasons outlined above. If the mark were treated merely as a separator, it would be impossible to constrain the location of the termination of the colon-expansion.

There is also an instance of a delimiting pair of marks where only the final mark is instantiated. This is in the case of the so-called *disambiguating* use of the comma. A comma is used to signal a clause boundary to avoid ambiguous or garden-path analyses. We find that the position of this comma coincides with the end a phrase that although not delimited conventionally with commas is adjunctive in nature. Therefore we can license the use of the final comma under the delimiting category to definitively mark the end of the preceding phrase. The initial mark for this delimiting pair is, as with the final mark of the colon-expansion, a null mark, to ensure that the interpretation given to the adjunct phrase is the un-delimited one rather than the delimited one.

Function

Delimiting punctuation marks are particularly useful to language processing, since they perform strong syntactic and semantic functions.

Syntactically, the delimiting marks signal the presence of an adjunct, and mark phrasal boundaries. Furthermore, in positions where it is possible that an adjunct is either restrictive or non-restrictive, delimiting punctuation marks it as being non-restrictive (and, strictly speaking, lack of punctuation marks the adjunct as restrictive). These adjuncts can come either before, after or in the middle of the phrases they modify, depending on the precise syntactic category of both adjunct and head-phrase.

Semantically, the precise variety of punctuation used to delimit the adjunct can indicate a degree of rhetorical balance, or relevance, of the delimited item to the rest of the text, and delimiting punctuation also specifies some discourse relation indicating that the delimited discourse fragment is a parenthetical to and an elaboration of the surrounding, higher-level discourse.

The degree of rhetorical balance is not so much an absolute measure as a relative one. If just one delimited fragment occurs in a text, it is impossible to tell how far removed it should be in terms of relevance from the main discourse structure. All that can be extracted, in terms of rhetorical balance, is that the delimited fragment should be at a less relevant level. However, if two or more delimited elements occur in the same discourse, then it is possible to establish relative degrees of aggregation (i.e. that some of the delimited elements are more closely related to the main discourse than others) via

a simple hierarchy of importance, or relevance to the main discourse (9.16).

(9.16) , ≫ — ≫ ( )

The colon-expansion does not function in the standard manner for delimiting punctuation, since syntactically it can contain the linguistic entities subcategorised by the preceding verb (and therefore does not really contain an adjunct), and semantically it tends to specify a richer set of relations than normal delimiting punctuation, and is unlikely to specify the parenthetical relations.

Of rather less use than all the other marks mentioned so far in this category are delimiting quotation marks. Semantically, they specify a great deal, usually narration or, in the case of scare-quotes, some unusual facet of the lexical items delimited by them (e.g. metaphor). Syntactically, however, they are almost useless. Although they can act in a similar manner to colon-expansions regarding verbal subcategorisation, it is equally possible for quotation marks to delimit a text fragment that is not a syntactic constituent category (9.17). This is backed up by Doran in her investigation (1996) where she states that quotation marks are not adequate for either identifying or constraining the syntax of quoted speech.

(9.17) According to FBI agent David ("Woody") Johnson, "a white male with an indistinguishable" American accent warned that a bomb would go off at the park within 30 minutes. (Gleick, 1996)

It should also be stated that, other than the obvious sentential-chunking role, the sentence-level delimiter pair (sentence-initial capitalisation and full-stop) have little function either syntactically or semantically, since syntactic analysis is invariably carried out below the sentential level, and the semantic relations that hold between adjacent sentences are too variant to be reliably indicated by punctuation.

- **Conjunctive Punctuation** — also described in (Nunberg, 1990) as *separating* punctuation.

Form

Unlike the paired adjunctive punctuation, conjunctive punctuation marks occur as singletons, independent of any other punctuation mark or phenomenon. Since they connect linguistic entities (lexical items, syntactic categories, etc.) they are not subject to absorption, although they are still able to absorb delimiting marks.

Marks that can function in this manner are the comma, the dash and the semi-colon.

Function

As the name *conjunctive punctuation* indicates, the function of these punctuation marks is to join together portions of adjacent text. Very often, the marks act as cataphors, resolving to an actual lexical coordination present between the final two items to be joined (any

conjunctive punctuation mark immediately preceding this mark serves only to properly segment the structure — it does not also resolve to the lexical conjunction itself) (9.18).

(9.18) a. We bought books, tapes, CD's and T-shirts.
b. Are we going to fly to Rome, drive to London, or walk to Edinburgh?
c. The members spoke to Mr Foot, a farmer; Ms Smith, a lecturer; and a social worker.

It is not necessary for there to be a final lexical coordination for the conjunctive punctuation marks to resolve to. In the absence of such a lexical item, the punctuation marks can also act exophorically, taking on the meaning of a weak conjunct (9.19) that can nevertheless be overridden should there be a suitable contextual or pragmatic reason (9.20).

(9.19) a. It was a tall, ugly, post-war municipal building. (Jarvie, 1992)
b. He was a distinguished academic geographer; he had wide-ranging research skills; he published books in all sorts of subject areas; he was also a devoted husband and father. (Jarvie, 1992)

(9.20) a. I will come on foot, you should drive.
b. We liked John; we disliked his politics (Jarvie, 1992).

As is clear from the sets of preceding examples, there is virtually no restriction on the syntactic entities that can be conjoined by punctuation marks from this category, from sentences to individual lexical items, such as adjectives. The only restriction is that the items to be conjoined be of a similar category (as is the case with ordinary, lexical coordinations).

Since there is no limit to the number of these punctuation marks that can occur in a single orthographic sentence, the potential for ambiguous analyses arises. This is resolved by the use of another power hierarchy for conjunctive punctuation marks which determines which marks can scope over others (9.21).

(9.21) ; ≫ — ≫ , ≥ **and** etc.

Semantically, there is little role for the marks from this category to play. The chunking of the text, which could be seen as a process of semantic or discourse aggregation, has already been performed in the syntactic analysis using the hierarchy in (9.21). If the punctuation marks are cataphorically resolved to a lexical coordination, then they take on the normal syntactic and semantic attributes of those lexical items. When the punctuation marks are exophorically resolved, however, and the weak conjunctive interpretation is applied, then if the elements that are conjoined by the punctuation are sentential (i.e complete discourses) the punctuation marks could be allowed to specify a weak Continuation relation. However, as with the weak conjunctive interpretation this relation would have to be over-rideable by context and pragmatics and even without this defeasability constraint it remains debatable whether the relation of continuance holds with any reliability at all between juxtaposed sentences. Therefore, it seems most advisable to disregard

any specific semantic contribution by conjunctive punctuation marks, and to treat them as functional only at the syntactic level (unless, as discussed, they resolve to lexical coordinations).

- **Stress Markers**

  The question mark, exclamation mark and ellipsis are best treated in a separate category from other punctuation since their form and function differ significantly from the other two categories we have explored.

  Although graphically similar to other marks of point punctuation, such as the full stop, they do not play a role in the standard conjoining or delimiting function that these marks play. Rather they graphically absorb (Nunberg, 1990) these marks, taking on their functions in the process, whilst still performing their own functions at the same time.

  The extent of these functions differs a great deal between the three marks. Their syntactic functions are almost negligible: any syntactic role they do play (marking the end of a clause or sentence, for example) is due to the syntactic function of the point punctuation mark they have absorbed. Almost the only true syntactic function of a stress marker is that particular use of the question mark that permits an alternate constituent order in the preceding clause (9.22).

  (9.22) *Is it possible.
  Is it possible?

  It is only really in terms of semantic function that the stress markers become important. As discussed in chapter 8, the ellipsis is best treated in orthography by lexicalisation to *etcetera*, some similar (possibly underspecified) lexical entity, or an interruption or pause in quoted speech. The exclamation mark allows for the preceding portion of the text to be emphasized in terms of importance to the overall discourse structure, strength of individual belief, or truth value. This emphasis could be manifested within the discourse structure either by use of particular rhetorical relations or by insertion of extra assertions into the analysis.

  The question-mark has possibly the richest contribution to make, by specifying that some portion of the immediately preceding discourse is a question, which admits several different treatments in discourse structure. These treatments, similar to the function of the exclamation mark, can either involve the specification of particular rhetorical relations or the addition to the discourse structure of unresolved, or anaphoric, terms. Some sentences that terminate with a question mark could be recognised as questions without the presence of the stress marker, due to the presence of some *wh*-word or phrase, but the presence of the mark helps to emphasise the questional nature of the sentence. Thus the information contributed by the stress marker in these cases serves to emphasise or confirm the discourse analysis generated from the preceding sentence alone. However, it is still necessary for the question mark to be associated with such specific semantic functions since there are sentences whose questionality is indicated solely by the final

stress marker (9.23). In these cases, therefore, the functionality of the question mark and the information it contributes are crucial to the success of the complete language processing task.

(9.23) John was right.
John was right?

9.1 Overall Hierarchy

Is is now possible to combine all the categorial distinctions and descriptions made in the previous sections into an overall hierarchy of punctuation, describing the nature of punctuation, its form and variety, and the various facets of its functionality. This hierarchy is as described in figure 9.1.

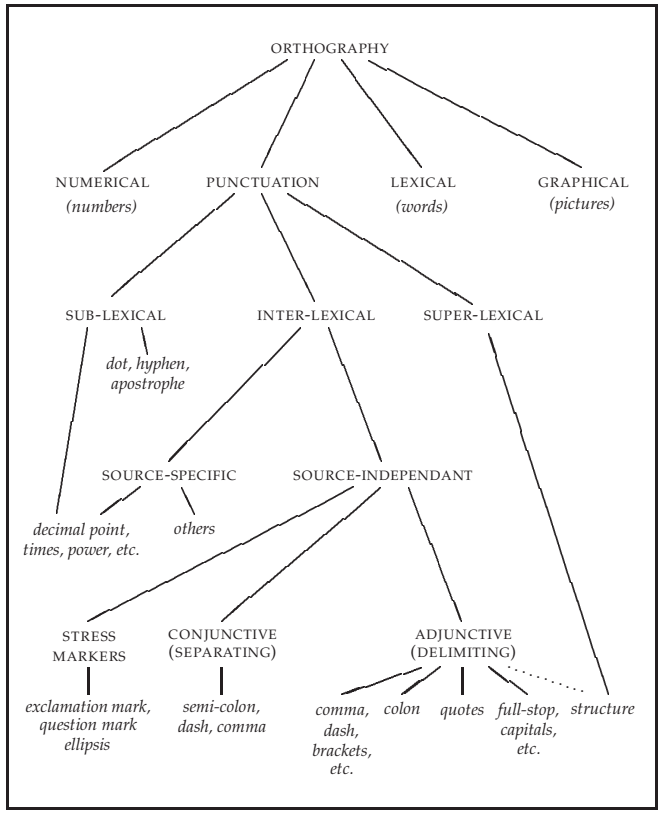


Figure 9.1: Taxonomical Hierarchy for Punctuation

ten The Theory of Punctuation

"It's all very well in practice, but will it work in theory?"
(Garret Fitzgerald, when prime minister of Ireland)

It is now possible to integrate all the work of the preceding chapters into a theory of punctuation that is coherent and suitable for computational implementation. However, before this is done the scope and purpose of the theory should be clearly defined: what this theory is *not* intended to be is a grammar or set of rules that can be directly inserted in any appropriate language processing system to implement punctuation functionality; rather the theory is intended as a set of guidelines and principles, which can be followed as closely as necessary or possible in the system concerned, to enable punctuation marks in the text to be processed appropriately.

10.1 Field of Applicability

The first decision point when using this theory is to determine whether the prime function of the language processing system it is to be used with is one of analysis or generation. The reason this distinction needs to be made so early lies in the non-specificity of punctuation marks. In analysis, a safe, Gricean assumption can be made that the material to be analysed is (reasonably) well-formed and contentful. Therefore all and any punctuation can be assumed to be important, and worthy of interpretation. This requires an inherent degree of flexibility and ambiguity in any system, since, in certain cases, different punctuation marks may be used interchangeably, and the same punctuation mark can have multiple functions. In generation, this degree of flexibility would be counter-productive. A far more prescriptive approach is needed in this field, to dictate the best, most clear, insertions of punctuation into the generated material.

Therefore, in the case of natural language generation, the majority of the current theory is not directly applicable. All that is needed is a set of prescriptive punctuation realisation rules,

as defined in (Nunberg, 1990) and discussed in section 3.2, the principles of punctuation interaction (such as point absorption), and some *pouring* rules (similarly defined and discussed) to determine the appearance of the text in the output medium. These are all relatively straightforward, and examples are discussed in (Meyer, 1987) and (Nunberg, 1990). The principles relating to the syntactic and semantic functions of punctuation in the theory presented below could also be relevant to the formation of the realisation rules. The reason why the insertion of punctuation into language generation systems has classically been problematical though lies not so much in the operation of the punctuation system as the difficulties in generating the appropriate lexical syntactic constituents from internal knowledge representations (e.g. complex clausal and sub-clausal structure).

Thus the theory presented here is designed primarily for analytical purposes, where flexibility and robustness are essential if the maximal amount of information possible is to be extracted from any punctuation marks present in the text.

10.2 Punctuation Determination

The first, most crucial, operation to be accomplished is to determine which the punctuation marks are in input text. As discussed in section 1.1, any non-alphanumeric character in the input can be assumed to be punctuation, with certain reservations. Usually punctuation, as applicable to current analysis techniques, occurs in word-external positions (i.e. inter-lexical punctuation), so care should be taken with any punctuation marks that are word- or number-internal. A morphological processing facility, or lexical look-up, should be able to isolate phenomena that are irrelevant to the field of punctuation, such as hyphenation, apostrophes, formation of more complex numerical entities and abbreviations. The work reported in chapter 5 then suggests that the remaining, true punctuation can be further subdivided into two categories: source-dependent, e.g. (10.1), and source-independent (10.2).

(10.1) **@ # $ % & ;-)**

(10.2) **. ! ? ... ; : — , “ ” ( )**

The source-independent category is invariant, and forms the bulk of the punctuation that is likely to be encountered. The only point to be aware of is the precise quotation and bracketing practices prevalent in the text(s) to be analysed. If only one variety is likely to be present, then the system can be designed to respond to just that variety — otherwise a more complicated arrangement will be necessary to recognise and interpret brackets and quotation marks appropriately during processing.

The source-dependent category, as the name suggests, depends on the source material for its precise constituency. Therefore pre-definition of this category is much more problematic. If the material to be analysed has been produced with the consistent use of particular stylistic conventions (a newspaper, for example) then the likelihood is that only a small number of 'special' punctuation marks will have to be licensed. If this is not the case, however, the best that can be done is to try to account for all the most common marks likely to occur in whatever genre the system is to apply to (e.g. currency symbols, mathematical symbols, %, ;-) in USENET).

10.3 Segmentation

Once the punctuation marks have been recognised and tagged as such, the text can be segmented appropriately for processing. This is a difficult but important task, which this thesis has not addressed, except peripherally during corpus processing in chapter 5. It is possible to use punctuation marks to subdivide input text into paragraphs and sentences, as in (Palmer and Hearst, 1994), but more remains to be done in this field. Systems so far have often relied on input text being pre-segmented manually, but as processing becomes larger scale (i.e. millions of sentences) this task will have to be performed reliably automatically. It is important not to confuse sentential boundaries which are internalised (e.g. inside a quotation) with external text-level sentence boundaries. While both are important from the point of view of parsing, it is essential to recognise the latter as distinct from the former to preserve the semantic and discursive structure.

Problems can arise with the confusion of abbreviative points and full-stops, especially when these occur at the end of clauses or sentences. In certain circumstances, additional examination of the sentential context and syntactic categories can assist the analysis (which necessitates a certain degree of parallelism between the segmentation and parsing proecsses). However, there are cases where it is impossible to segment reliably without a great deal of semantic, pragmatic and world knowledge, for example (10.3) where there could be two sentences with a sentence boundary before *Friday* or the whole example could be a single sentence with a constituent boundary after *Friday*.

(10.3) Build the drum-kit by 5 p.m. Friday and Saturday will be a nightmare for me.

10.4 Syntactic Function

The relationship of inter-lexical punctuation marks to syntactic functionality is best described by first dividing the punctuation phenomena into two different categories. The first, broadly equivalent to Nunberg's (1990) category of *delimiting* punctuation, is **adjunctive** punctuation; the second, sharing some similarity with Nunberg's *separating* punctuation, is **conjunctive** punctuation.

Before investigating these, however it is worth mentioning the treatment of the source-dependent punctuation marks. These should all be lexicalised, if possible, so that they take on the syntactic function of the words they represent (e.g. % ⇒ *percent*). Some of the marks, however, have coordinating functionality (mathematical symbols, &, etc.) and therefore there is an additional syntactic dimension to these. With these co-ordinating marks, too, it is necessary to maintain some of the original punctuation sense so that correct grouping can be performed (so that, for example, & is not confused with *and* in (10.4), and so that mathematical formulae (if parsed) are grouped correctly).

(10.4) I like fish & chips and sausage & egg.

Conjunctive Punctuation

The punctuation marks that fall into this category are most frequently the semi-colon and the comma. It is possible, in addition, for certain uses of the dash to also fall into this category. What the marks of this category have in common is that they all come between similar syntactic categories (for example, between lexical sentences, or noun phrases.

While the comma can occur between almost any syntactic categories (10.5), the semi-colon usually only occurs between lexical sentences (10.6). When it occurs other than at sentential level, Nunberg's (1990) principle of semicolon-promotion (c.f. section 3.2) is in effect, and the restrictions of that principle apply (i.e. semi-colons can only occur at the top level of the clausal structure, final lexical coordination must be preceded by a semi-colon, etc.), as in (10.7). The dash, when used, can only occur once (thus a maximum of two categories can be conjoined), should separate sentential clauses, and is ideally followed by a lexical coordination (10.8), although this is not always necessary. Similarly, comma- and semicolon-conjoined lists do not always need a lexical coordination between the final elements, although comma-conjoined lists usually have one. An alternative method of treating the sort of dash encountered here is by regarding it as a dash-expansion (similar to colon-expansion) (White, 1995). However, under the current analysis, that would not only place both uses of the dash in the same procedural category, but also lead to considerable confusion between the form and functionality of the dash-expansion and the dash-interpolation.

(10.5) I enjoy apples, pears, mangos and cherries.

(10.6) I'll take the high road; you take the low road.

(10.7) At the meeting were John, a teacher; Mary, a solicitor; and Edward, a policeman.

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

These principles regulate the possible positions of occurrence of these punctuation marks; their syntax is inherited cataphorically from any lexical coordination present (usually immediately preceding the final item, but not necessarily). Thus the list in sentence (10.5) is interpreted as *apples and pears and mangos and cherries*. If the sentence had the form of (10.9), the same meaning would be present. The punctuation marks are not lexicalised completely, however, since they still have a structural role to play. Thus in (10.10), although the commas inherit the syntax of *and*, they remain a 'higher-level' co-ordination than the *and*s within the listed items. If there is no final lexical conjunction for the punctuation marks to inherit their syntax from (10.6), their syntax defaults to that of *and*. Note that because the inheritance should be reasonably formal, sentences with contradicting co-ordinations are blocked for linear interpretation (10.11).

(10.9) I enjoy apples, and pears, mangos, cherries.

(10.10) I enjoy fish and chips, bread and butter, and tea and biscuits.

(10.11) *I would like cheese, and grapes, or fish.

These principles can therefore be summarised in a set of rules (with appropriate restrictions to prevent attachment ambiguities, such as forced right-branching).

$\mathcal{X} \Rightarrow \mathcal{X}$ { **; (coord)** $\mathcal{X}$ }$^{+}$
$\mathcal{Y} \Rightarrow \mathcal{Y}$ { **, (coord)** $\mathcal{Y}$ }$^{+}$
$\mathcal{Z} \Rightarrow \mathcal{Z}$ **— coord** $\mathcal{Z}$

Adjunctive Punctuation

The punctuation marks used in this category always come in pairs. They are comma—comma, dash—dash, left-bracket—right-bracket, capitalisation—full-stop and colon—null-element. However, these pairs do not always appear together at the surface-orthographic level due to the operation of a modified version of Nunberg's *absorption* principles (c.f. section 3.2). Either the initial or the final mark can be absorbed according to the hierarchies illustrated in the previous chapter in examples (9.13) and (9.14). It is always the case, however, that at least one of the initial or final marks is present to signal the punctuation feature.

These absorption phenomena can be treated in the grammar in a variety of ways. Formal use of them would introduce procedurality into the grammar and require the positing of underlying forms that the text does not attest. It is neater to implicitly take notice of absorption phenomena in the design of the grammar, and cover their effects by the use of extra features, such as the cliticising and feature-passing in (Jones, 1994b).

The commonality between punctuation marks in this category is that they all signal, and delimit, adjunctive structures (i.e. non-key, modifying syntactic constituents). The rule for the usage of this punctuation is:

> **when these punctuation marks are attached to a syntactic entity, the resulting entity must be a non-head daughter in a syntactic structure where head-daughter and mother entities are of the same syntactic category.**

Thus these adjunctive punctuation-delimited phrases can occur either before or after the syntactic entity they are modifying, with appropriate absorption taking place. The punctuation marks themselves contribute little concrete syntactic information to the parse, other than as rudimentary indicators of phrasal boundaries, unless the syntactic analysis is sophisticated enough to be able to distinguish between entities such as restrictive and non-restrictive relative clauses (marked by comma-delimitation), in which case this information can be included in the parse.

The rules to implement adjunctive punctuation would take the following form, with the additional provision of attachment restrictions and appropriate feature-passing conventions to implement not only absorption phenomena but also the restrictions on the content and position of certain adjuncts (such as the colon-expansion, as discussed in section 3.2). Rules expressed in greater detail (explaining permissible categories for the various marks) can be found in chapter 6.

$\mathcal{A} \Rightarrow \mathcal{A}$ **:** $\mathcal{K}$

$\mathcal{B} \Rightarrow \mathcal{B}$ **'('** $\mathcal{L}$ **')'**
$\mathcal{C} \Rightarrow \mathcal{C}$ **,** $\mathcal{M}$ **(,)**
$\mathcal{C} \Rightarrow$ **(,)** $\mathcal{M}$ **,** $\mathcal{C}$
$\mathcal{D} \Rightarrow \mathcal{D}$ **—** $\mathcal{N}$ **(—)**
TOP $\Rightarrow \ll$ **cap** $\gg \mathcal{S}$ **.**

Other Punctuation

The two categories above do not account for all the inter-lexical punctuation likely to occur in text. The quotation marks should really have been included under adjunctive punctuation, but they do not have such a clear-cut role. In (10.12), for example, the quoted material is actually subcategorised for by the verb *declared*. In addition, the content of a quoted phrase is not always a complete syntactic constituent (10.13), causing problems to a possible syntactic analysis. Therefore it is best to either give quoting punctuation a distinct syntactic treatment from normal adjunctive punctuation, or preferably to ignore it from the point of view of syntax and treat it as a purely semantic phenomenon, which would make it easier to cope with the phenomenon of quote transposition. This would only require appropriate subcategorisation information being available for the relevant verbs used with quoted speech. Further special consideration is needed for the treatment of other punctuation marks surrounding quotation (10.14). Rather than declare yet another use for the comma or the colon, and stipulate their syntactic removal in the context of quotation, it would make most sense to treat them in their normal adjunctive fashions. The quoted material thus becomes adjunctive/delimiting, which can be exploited by suitable subcategorisation of the verbs involved.

In fact, Doran finds in her investigation of quotation phenomena (1996) that quotation marks cannot reliably specify or constrain the syntax of the quoted phrase, and that using other surrounding punctuation or the relevant verbs of speech is more productive. Thus it is best for quotation marks to be treated as non-syntactic phenomena, but with a semantic contribution, as discussed later.

(10.12) He declared "It's an outrage."

(10.13) According to FBI agent David ("Woody") Johnson, "a white male with an indistinguishable" American accent warned that a bomb would go off at the park within 30 minutes. (Gleick, 1996)

(10.14) "The time has come," he said, "to talk of many things."

The second class of punctuation marks that have been ignored thus far are the stress markers, which cannot really be absorbed into either conjunctive or adjunctive categories. It is best to treat these marks as distinct from either category, and bearing no explicit syntactic information (just the semantic, stress information). The only case where these marks interact with our syntactic functionality of punctuation is when they absorb a sentence-final full-stop. In this case, they inherit the syntactic features of that full-stop, which (since absorption should be implemented implicitly in the grammar) means a redefinition of the TOP rule including each

stress marker. The stress markers do not perform any absorption activities with other point punctuation marks.

10.5 Semantic Function

As described in chapter 8, the semantic functionality of punctuation marks is best described in three categories: null functionality, lexical functionality and discourse functionality (which is then further subdivided).

Null Functionality

The punctuation marks that have no semantic significance are final adjunctive markers, conjunctive marks that occur immediately befoe lexical co-ordinations, and the sentence-initial capitalisation feature.

In semantic analysis, it is only necessary to indicate the subordinate nature of an adjunctive-delimited phrase, which the initial adjunctive mark will have done (possibly via explicitly marking the subordinate or non-restrictive nature of the phrase in the output of the syntactic analysis). The syntactic analysis will capture the extent of the adjunctive phrase, so there is no need for the explicit signalling of a re-entrancy into the higher-level discourse. Thus the only role for the adjunct-final marker is in the syntactic analysis, to indicate the completion of the adjunctive phrase. This even holds true in the cases where it is the adjunct-initial marker that has been absorbed by a higher level phenomenon. The adjunct-final marker signals the presence of the adjunctive phrase in the syntactic analysis, and then the syntactic analysis must signal the nature of that adjunct to the semantic processing component. Note that this may involve enhancing the detail and functionality of the syntactic analysis.

Conjunctive punctuation marks that occur immediately before lexical coordinations can also be treated as semantically void. All that they would do anyway is to resolve to that coordination, which is already present, rendering them redundant. They do fulfil an important role, of course, in determining the aggregation of the text, but this again is something that should have been inherent in the output of the syntactic analysis stage, which is why the marks can be ignored at this semantic stage. Of course, there is a specialised case where this does not apply (10.15), but the unique nature of this example should be discovered in the syntactic analysis stage, where it should be clear that the final coordination cannot be functioning in a coordinating role, since there is nothing following it.

(10.15) There are four words in the English that I love: except, or, and, but.

As discussed in chapter 8, although the sentence-initial capitalisation should behave in a similar manner to other opening delimiting marks, it is almost impossible to determine what the semantic functionality of the feature might be since there are no semantic functions that hold reliably between adjacent sentences. Therefore it seems more appropriate to declare this feature as semantically non-functional than to try to extract some vastly under-specified connective relation from it.

Lexical Functionality

As in the case of syntactic functionality, all those marks that are lexicalised also inherit the semantic functionality of their lexical entries. These therefore include all the various source-dependent punctuation marks, as well as some of the source-independent ones. The ellipsis (…), for example, can be semantically lexicalised as *etcetera*, as an underspecified entity or in the case of quoted speech as interruption or pausing.

The most common category of semantically lexicalised marks, however, is the conjunctive punctuation marks. The commas and semi-colons (not dashes, since as described above, they can only occur before lexical coordinations, and are therefore declared semantically empty) that act to conjoin linguistic items are lexicalised cataphorically to the lexical coordination in the list, if present (as discussed in section 10.4). If a lexical coordination is not present, the marks are lexicalised exophorically to a very weak version of the semantics of *and*. The reason for this is that a conjunctive interpretation is not always the most appropriate — in some cases a contrastive one is preferable. Therefore a weak conjunctive interpretation is less likely to disrupt the sense of the material.

Discursive Functionality

The discursive functions of rhetorical balance and aggregation, which Dale (1991) attributes to punctuation, have been discussed in chapter 8 and considered more appropriate as a stage of the syntactic function of the punctuation. The single exception to this is the relative degree of rhetorical balance indicated by the choice of punctuation for an adjunctive-delimited item. Thus a comma-delimited phrase is closer, rhetorically, to the surrounding discourse than a dash-delimited phrase, which in turn is closer than a bracket-delimited phrase.

The main portion of discursive functionality, however, relates to specification of discourse relations, and mainly to adjunctive punctuation marks. The parenthetical adjunctive punctuation marks (comma, dash, breacket) can indicate both Comment and Elaboration relations holding between the delimited discourse fragment and the main discourse outside it, as discussed in chapter 8. The colon-expansion, traditionally the most semantically rich of all normal point punctuation marks, contributes a defeasible default relation of Elaboration. This should be overrideable, however, through context, by either a set of possible relations or a different relation altogether. Since this relates to the field of context and pragmatics, however, the exact mechanism of such a specification lies outside the realms of this investigation.

The remaining punctuation marks — stress markers and quotation marks — are semantically specified as laid out in chapter 8. The exclamation mark specifies emphasis (in terms of relevance or importance to the discourse) of the truth value of a proposition. The question mark places uncertainty on the truth value of a discourse structure, or alternatively acts together with specific questioning lexical items to place some other uncertainty into the discourse structure.

Quotation marks, specified as syntactically non-functional, function most obviously to trigger the Narrative discourse relation. Scare quotes should be handled in some system-specific manner, either by relaxing semantic and pragmatic constraints on the lexical items they surround (thus admitting 'unusual' terms into an interpretation) or by treating the quoted

items as metaphoric.

10.6 Pragmatic Function

There is little contribution of punctuation to the pragmatics of language, other than those contributions that are passed along by syntactic or pragmatic analysis. The only immediately apparent case of pragmatic function is in the case of multiple stress markers (a practice that is, of course, frowned upon in most style guides). The usage in (10.16), for example, might represent a progressive increase in volume or desperation (Parkes, 1992), whereas in (10.17) it cannot refer to volume but could refer to desparation, concentration or emphasis.

(10.16) Stop! Stop!! Stop!!!

(10.17) "Stop!!" she whispered.

10.7 Summary

In this chapter I have presented a theory, consisting of general rules and principles, which should enable a treatment of punctuation to be inserted into a computational language processing system. It is not intended to be an exhaustive and precise description that can simply be inserted as a unit into an arbitrary analysis package; rather it is a schema which provides a guide for the insertion into a processing system of sufficient functionality to enable information to be extracted from the punctuation present in the text to be analysed.

There is some confusion between the semantic and syntactic roles that punctuation marks are able to play, especially regarding roles such as aggregation and rhetorical balance. In some cases, a unified approach such as that of Say and Akman (1996) which investigates punctuation marks for their total *information* content, are appropriate, but it is difficult to see how such unified approaches would succeed on a general implementational level. Therefore the current theory has been separated, as far as is possible and useful, into syntactic and semantic components.

An additional point that should be emphasised about punctuation is that its usage is very frequently idiosyncratic. It is far more likely for punctuation marks to be used in unanticipated situations than it is for lexical items to be used in unusual orders and positions, a situation that is encountered relatively infrequently. Despite this, however, approaches have been proposed for the treatment of language that does not correspond to the normal principles of syntax, such as (Jones, 1995a). Therefore, especially with the advent of robust, flexible analysis methods for normal language analysis, the flexibility of a punctuation analysis system should be very high, so that marks that occur in highly unusual places do not cause the analyis to fail.

eleven Multilingual Applicability of the Theory

I don't like people who punctuate, especially people who do that <quotes with fingers>. I hate that. "Oh yeah, he's a <"> comedian <"> if you get my drift." Don't you get the overwhelming desire to grab their fingers and [crush them]? If somebody's doing it to you, try the thing that I do. If somebody's doing it to me, I start punctuating everything. I do commas, full stops, the whole lot. I say, "You know <draws comma> I was just saying that to my wife only yesterday <draws full-stop> <quotes> Darling <comma> I said <full-stop> She said <quotes> What <draws question mark> I said <comma> <quotes> Oh shut up <draws exclamation mark>"

(Billy Connolly, "World Tour of Scotland")

Systems for language processing are placing more and more emphasis on multilingual capabilities, so it would be interesting to see how easily the theory postulated in the previous chapter will transfer to other languages. In its current form, of course, it is rather unlikely that the theory could be immediately applied to languages other than English (other forms of English maybe, but not languages separate from it). However, if the punctuation systems of another language differ only from that of English in a few circumstances, it is possible that the theory could be restated in the form of a punctuation 'engine' and a set of language-specific rules and principles.

11.1 Western Languages

As described in chapter 2, the current punctuation system we use stems from that used by the Romans and, later, in the Holy Roman Empire. Therefore, since the writing traditions of

other European countries are based on the same historical systems, the principles of punctuation should not be that different from one-another. Indeed, the systems of German and French, to take two West European examples, are almost identical to that of English; the most notable exception being a greater degree of prescriptivity in punctuation usage in particular situations (to take a frequently-mentioned example, in German a comma is mandatory before all finite subordinate clauses). Other than that, the differences in the systems are almost trivial: point absorption, for example, seems to work differently in French (11.1) in that a dash does not absorb a comma, whereas semi-colon promotion works identically (11.2).

(11.1) Un travail plus sûr, plus qualifié — donc mieux rémunéré —, des perspectives… *(Le Figaro)*

(11.2) …2003: vaccin contre le sida; 2006: guérison du sida; 2013: médicaments pour prévenir le cancer… *(Le Figaro)*

Indeed, so similar are the systems of punctuation that in investigations of the French punctuation system with respect to comma placement, Simard (1993; 1996) uses the punctuation system proposed for (American) English by Nunberg (1990) and claims it is equally applicable to French.

Examining some German style guides (in the non-editorial sense) a similar situation to that of French emerges (Berger, 1982; Baudusch, 1984). Usage of the various inter-lexical punctuation marks, and their variety, is broadly comparable to English. The only major exception for German is that the usage of the comma is rather more complicated, with particular situations being licensed or forbidden depending on the lexical entities to either side. Once again, some of Nunberg's absorption principles do not seem to hold, in particular regarding the dash (11.3–11.5).

(11.3) Das Mädchen, mit dem er sich verlobt hatte — übrigens gegen den Willen ihrer Eltern —, ist ihm schließlich doch untreu geworden. (Berger, 1982)

(11.4) Verächtlich rief er ihm zu — er wandte kaum den Kopf dabei —: „Was willst du hier?" (Berger, 1982)

(11.5) Zuletzt tat er das, woran niemand gedacht hatte — er beging Selbstmord. (Berger, 1982)

Sentence (11.4) illustrates the most noticable differences between punctuation use in various languages, namely the variant quotation marks. As illustrated, German uses an alternation of low and high signs (and note the shape of the signs is reversed from that used in English), and French usage favours so-called *guillemets* (11.6). Spanish quotation also prefers these marks (11.7) although it should be noted that these languages all make use of 'English' quotation marks for scare quotes.

(11.6) ≪oui, je n'y ai jamais pen…pen…pen…pensé≫ (Balzac, 1833)

(11.7) ≪antes de inventar signos nuevos básicos […] habría que explorar a fondo las posibilidades de los existentes.≫ (Poyatos, 1994)

The most obvious difference in Spanish punctuation, other than the quotation marks, is the delimiting nature of the stress markers. These were established in 1739 in a meeting of the Real Academia Española in Madrid as part of the some general rules for Spanish orthography (Parkes, 1992; Serrano, 1982), where to solve the problem of ambiguity relating to which part of the sentence was to be 'stressed' an initial, inverted version of the marker was proposed (11.8). This system has a great advantage over the non-delimiting system, in that it is possible to precisely denote the portion of the text to be questioned or emphasized (and so the marks can be used to stress just a single word, or phrase). Obviously, this delimiting system would be easy to incorporate into the theory formulated previously, since the stress markers could simply be re-cast as adjunctive markers.

(11.8) ≪¡Hombre, Goro, hoy…!≫
≪¿Quieres que entre el señor cura para reconciliarte?≫ (Poyatos, 1994)

Thus it seems that the theory proposed in this thesis would be suitable for multilingual usage, requiring a minimum of modification to operate in other languages. Some rules would need to be modified (for example relating to comma-placement in German) and so would the principles of punctuation-interaction (e.g. point-absorption), but in general, in terms of variety and broad usage and function, punctuation is similar in many Western languages.

There are other considerations of multi-lingual punctuation use that are of less relevance to the current theory, since they involve punctuation other than inter-lexical. In the case of sub-lexical punctuation, for example, it is common practice in Dutch to use the apostrophe to mark **any** plural term (so that the mark comes before the pluralising morpheme). This is easily dealt with in Dutch processing systems by formulation of a different account of apostrophe use to that of English. These procedural considerations should not impinge on the operation or use of the theory for inter-lexical punctuation marks, however.

11.2 Other Languages

It is more questionable that the theory would work with other, non-western languages that do not have the same broad orthographic basis as English. Indeed, many non-european languages have classically done without punctuation (Persian, Chinese, etc.) until the recent era of internationalisation, where western text is becoming commonplace world-wide and hence some of its traditions are being adopted into other text styles, to help with problems of ambiguity. In any case, even without a common historical base, it is still possible that the broad function of the punctuation systems will be similar.

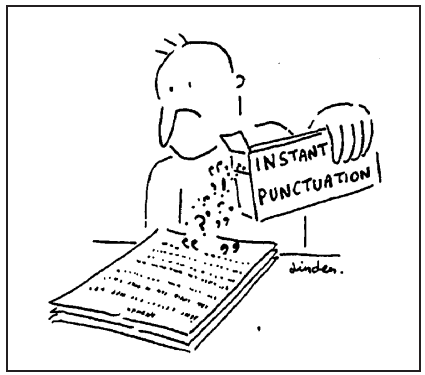
As Fornell reports (1996), Japanese has a number of punctuation characters that correspond closely to the English ones. The *maru* 。 has a function similar to the full-stop; the *ten* 、 functions as a conjoining or adjoining comma (although its usage is based more on prosodic than syntactic principles); and the *kagi* 「 」 function as quotation marks. Further to these marks, some 'English' marks are used directly, such as parentheses, dashes and a six-dot ellipsis. In addition, however, there are two specifically Japanese punctuation characters that do not have direct graphical correspondents in English: the *namigata* 〜 functions as a sort of dash

denoting a *from-to* relationship, and the *nakaten* ・ functions as a multi-purpose separator which sometimes corresponds to a slash in English (e.g. and/or), sometimes just to a space, and sometimes to separate foreign personal names.

It would not be too hard to integrate this punctuation system into the current theory, except for the specifically Japanese characters, which could be dealt with by assigning the *namigata* a lexicalised/syntactic function, and the *nakaten* a segmentation or lexical function, depending on its usage.

This integration of the theory with non-western languages is not certain to work in all cases, and with all marks, but as illustrated in certain cases it is possible to extend the scope of the theory to languages with non-western orthographies and with non-standard punctuation systems. This is a flexibility that is not only encouraging, but would be unlikely if the theory had been stated in a more concrete fashion, as a set of rules to be inserted into an arbitrary grammar.

twelve Conclusion



(Private Eye, 1995)

Employing a variety of techniques, a variety of facets of punctuation have been explored in this thesis: the variety of marks present in text, the syntactic functionality of the punctuation marks (both observational and theoretical) and the semantic and discourse roles that the punctuation marks have to play. These facets have been integrated into a theory of punctuation that does more than merely describe the use of individual punctuation marks: rather it describes the function of the punctuation system, and does so in a manner that is computationally applicable and tractable.

The theory is presented and designed in such a manner that it is not an inflexible, concrete set of rules which are simply to be inserted into an appropriate grammar. Instead it is a set of principles and concepts that can be followed and implemented to create punctuation functionality in arbitrary language processing systems. The syntactic and semantic portions of this functionality can be applied in the system separately or together.

It appears that the theory presented here is flexible enough to have potential multi-lingual application with a minimum degree of modification for most western languages, and even shows promise for languages based on a different orthographic system.

The theory draws upon notions of punctuation mark use presented in many style-guides and other work. In particular, however, it draws upon the system developed to account for the linguistics of punctuation by Nunberg (1990). It rejects Nunberg's proposition of a totally separate text-grammar, but combines certain elements of the text grammar with another valuable contribution of Nunberg's work, namely the punctuation absorption and transposition principles, and develops these into a rich account of syntactic and semantic punctuation functionality. In addition, practical distinctions are made in the variety of punctuation marks that are likely to be encountered, with suggestions for the most suitable treatments for the various categories.

Small-scale tests of this theory indicate its potential and functionality in the natural language processing environment. Since the exact implementational details do not form part of the specification of the theory, no particular claims are made for the precise implementational methods used in these tests. Usage of different formalisms in the addition of punctuation functionality to any system should not provide any significant problems, and may highlight areas where the theory could be further improved.

One of the most exciting potentials for punctuation is in conjunction with robust and partial parsing, and if punctuation can be successfully used to mark certain phrasal boundaries (as suggested in (Shiuan and Ann, 1996)) then the parsing task would be greatly simplified, especially for the longer sentences that are prevalent in much real-world text. Additionally, since there is some confusion as to the interferance of syntactic and semantic roles of punctuation (for example to indicate aggregation and rhetorical balance), any approaches that treat the subject in a unified manner (such as (Say and Akman, 1996)) will prove interesting.

The current work should integrate well with such systems, and will hopefully prove to be sufficiently flexible to enable modern language processing systems to be able to extract as much information from the graphical manifestations of their textual materials as they are able to extract from the linguistic material they contain.

one Bibliography

Scott Adams. 1992. *Build a Better Life by Stealing Office Supplies: Dogbert's Big Book of Business*. United Media, page 92.

Scott Adams. 1996. *"Dilbert" syndicated cartoon strip*. United Media Inc.

Paul Allardyce. 1884. *"Stops" or How to Punctuate*. Fisher Unwin: London, UK.

Hiyan Alshawi (editor). 1992. *The Core Language Engine*. MIT Press: Boston, Massachusetts.

Anonymous author. 1680. *A Treatise of Stops, Points, or Pauses* From the British Museum, reproduced as No. 65 of a Collection of Facsimile Reprints of *English Linguistics 1500–1800*.Scolar Press: Menston, UK.

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

Honoré Balzac. 1833. *Eugénie Grandet*. Cited in (Poyatos, 1994).

Lord Balfour . Cited in (Bennett, 1994) page 321.

Renate Baudusch. 1984. *Punkt, Punkt, Komma, Strich: Regeln und Zweifelsfälle der deutschen Zeichensetzung*. VEB Bibliographisches Institut: Leipzig, Germany.

Alan Bennett. 1994. *Writing Home*. Faber and Faber: London, UK.

Dieter Berger. 1982. *DUDEN: Komma, Punkt und alle anderen Satzzeichen*. Bibliographisches Institut: Mannheim, Germany.

B Bischoff. 1981a. Panorama der Handschriftenüberlieferung aus der Zeit Karls des Grossen. In (Bischoff, 1981b), pages 5–38.

B Bischoff. 1981b. *Mittelalterliche Studien*, **iii**. Stuttgart, Germany.

P Bohn. 1887. Das liturgische Recitativ und dessen Bezeichnung in den liturgischen Büchern des Mittelalters. *Monatshefte für Musikgeschichte*, **xix**.

Matthew Bond. 1996. Confessions of the academically challenged — Television Review. In *The Times* Newspaper, January 25th edition, page 43. News International, London.

Victor Borge. 1980. "Phonetic Punctuation". In *More Fun at One*, BBC Enterprises, Cassette no. ZCF399.

Edward Briscoe. 1994. Parsing (with) Punctuation. Rank Xerox Research Centre Grenoble, Technical Report MLTT-TR007.

Edward Briscoe. 1996. The Syntax and Semantics of Punctuation and its Use in Interpretation. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 1–7, Santa Cruz, California, June — also available as (Jones, 1996c).

Edward Briscoe and John Carroll. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels. In *Proceedings of the ACL/SIGPARSE 4th International Workshop on Parsing Technologies*, pages 48–58, Prague, Czech Republic.

Edward Briscoe, Claire Grover, Bran Boguraev and John Carroll. 1987. A Formalism and Environment for the Development of a Large Grammar of English. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 703–708, Milan, Italy.

Edward Briscoe and Nick Waegner. 1992. Robust Stochastic Parsing Using the Inside-Outside Algorithm. In *Proceedings of the AAAI Workshop on Statistically-based NLP Techniques*, San Jose, California.

T Brown. 1982. The Irish element in the Insular system of scripts to circa A.D. 850. In (Löwe, 1982), pages 101–119.

Bill Bryson. 1990. *The Lost Continent: Travels in Small Town America*. Abacus: London, UK.

D Bullough. 1973. *The Age of Charlemagne*. London, UK.

George Carey. 1939. *Mind the Stop: 1st Edition*. Cambrige University Press: London, UK.

George Carey. 1958. *Mind the Stop: 2nd Edition*. Cambrige University Press: London, UK.

John Carroll, Edward Briscoe and Claire Grover. 1991. A Development Environment for Large Natural Language Grammars. Technical Report 233, Cambridge University Computer Laboratory.

Lionel Casson. 1988. howandwhypunctuationevercametobeinvented. *Smithsonian*, October 1988, page 216.

Billy Connolly. 1994. *Billy Connolly's World Tour of Scotland*. Television Series: British Broadcasting Corporation.

Frank Cross. 1989. The Invention and Development of the Alphabet. In (Senner, 1989b).

Robert Dale. 1990. A Rule-based Approach to Computer-assisted Copy Editing. *Computer Assisted Language Learning*, **2**, pages 59–67.

Robert Dale. 1991. Exploring the Role of Punctuation in the Signalling of Discourse Structure. In *Proceedings of the Workshop on Text Representation and Domain Modelling*, pages 110–120, Technical University Berlin.

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

David Diringer. 1962. *Writing*. Ancient Peoples and Places, 25. Thames and Hudson: London, UK.

Christine Doran. 1996. Punctuation in Quoted Speech. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 9–18, Santa Cruz, California, June — also available as (Jones, 1996c).

Shona Douglas and Robert Dale. 1992. Towards Robust PATR. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France.

Shona Douglas and Matthew Hurst. 1996. Layout and Language: lists and tables in technical documents. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 19–24, Santa Cruz, California, June — also available as (Jones, 1996c).

Roddy Doyle. 1988 *The Commitments*. William Heinemann: London, UK.

Garret Fitzgerald. attributed to him, when prime minister of Ireland, at a meeting of the Irish Cabinet. In (Leaver, 1992).

J Fontaine. 1959. *Isidore de Seville et la culture classique dans l'Espagne wisigothique* Paris, France.

Jan Fornell. 1996. Punctuation in the Bravice English-to-Japanese Machine Translation System. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 25–33, Santa Cruz, California, June — also available as (Jones, 1996c).

H W Fowler and F G Fowler. 1930. *The King's English*. Wordsworth Reference (1993 edition): Ware, UK.

Werner Frey and Hans Kamp. 1987. Plural Anaphora and Plural Determiners. Unpublished Manuscript.

I J Gelb. 1963. *A Study of Writing: The Foundations of Grammatology*. Revised Edition. University of Chicago Press: Chicago, IL.

Henry Gleason. 1965. *Linguistics and English Grammar*. Rinehart and Winston, New York.

Elizabeth Gleick. 1996. Terror's Venue. In *Time* Magazine, August 5th edition, page 13. Time International: Amsterdam, Netherlands.

Kenneth Grahame. 1859–1932. Cited in (Bennett, 1994) page 225.

M W Green. 1989. Early Cuneiform. In (Senner, 1989b).

Sidney Greenbaum and Randolph Quirk. 1990. *A Student's Grammar of the English Language*. Longman Group: Harlow, UK.

Barbara Grosz and Candy Sidner. 1987. Attention, Intentions and the Structure of Discourse. *Computational Linguistics* **12(3)**, pages 175–204.

Patrick Hanks, William McLeod and Laurence Urdang. 1986. *Collins English Dictionary*. Collins: London, UK.

L Hector. 1966. *The Handwriting of English Documents*. London, UK.

Catherine Hilton and Margaret Hyder. 1992. *Getting to Grips with Punctuation and Grammar*. Letts Educational: London, UK.

Jerry Hobbs. 1985. On the Coherence and Structure of Discourse. CSLI Report No. CSLI-85-37: Stanford, California.

Jerry Hobbs. 1991. SRI International: Description of the TACITUS System as used for MUC-3. In *Proceedings of the Third Message Understanding Conference (MUC-3)*, DARPA, pages 200–206

L Holtz. 1981. *Donat et la tradition de l'enseignement grammaticale: Etude sur l'**Ars Donati** et sa diffusion (iv^e–ix^e siecle)*. Paris, France.

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

Eduard Hovy and Yigal Arens. 1991. Automatic Generation of Formatted Text. In *Proceedings of AAAI-91*, pages 92–97.

R Lee Humphreys. 1993. Book Review of "Geoffrey Nunberg: The Linguistics of Punctuation". *Machine Translation* **7**, pages 199–201. Kluwer Academic Publishers: Netherlands.

T F Husband and M F A Husband. 1905. *Punctuation: Its Principle and Practice*. Routledge and Sons: London, UK.

Ray Jackendoff. 1977. *X-bar Syntax: A Study of Phrase Structure*. MIT Press: Cambridge, Massachusetts.

Gordon Jarvie. 1992. *Chambers Punctuation Guide*. W & R Chambers: Edinburgh, UK.

Michael Johnston, Branimir Boguraev and James Pustejovsky. 1995. The Acquisition and Interpretation of Complex Nominals. In *Working Notes of the AAAI Spring Symposium on the Representation and Acquisition of Lexical Knowledge*, Stanford University, Palo Alto, California.

Michael Johnston and Federica Busa. 1996. Qualia Structure and the Compositional Interpretation of Compounds. In *Proceedings of the ACL/SIGLEX Workshop in the Breadth and Depth of Semantic Lexicons*, pages 77–88, Santa Cruz California, June.

Bernard Jones. 1994a. Can Punctuation Help Parsing? Esprit Acquilex-II Working Paper No. 29, Cambridge University Computer Laboratory, UK.

Bernard Jones. 1994b. Exploring the Role of Punctuation in Parsing Real Text. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 421–425, Kyoto, Japan, August.

Bernard Jones. 1995a. Is there a Role for Parsing in Natural Language Processing? In *Proceedings of the 4th International Conference on the Cognitive Science of Natural Language Processing*, Dublin, Ireland, July.

Bernard Jones. 1995b. Predicating Nominal Compounds. In *Proceedings of the 17th Annual Cognitive Science Conference*, pages 130–135, Pittsburgh, Pennsylvania, July.

Bernard Jones. 1995c. Exploring the Variety and Use of Punctuation. In *Proceedings of the 17th Annual Cognitive Science Conference*, pages 619–624, Pittsburgh, Pennsylvania, July.

Bernard Jones. 1995d. Nominal Compounds and Lexical Rules. In *Proceedings of the* ACQUILEX *Workshop on Lexical Rules*, Cambridge, UK, August.

Bernard Jones. 1996a. Towards Testing the Syntax of Punctuation. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96), Student Session*, Santa Cruz, California, June.

Bernard Jones. 1996b. Towards a Syntactic Account of Punctuation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark, August.

Bernard Jones (editor). 1996c. *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*. Technical Report HCRC/WP2, Human Communications Research Centre, University of Edinburgh, UK.

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

Alex Lascarides, Edward Briscoe, Nicholas Asher and Ann Copestake. 1996 Persistent Order Independant Typed Default Unification. *Linguistics and Philosophy* **19:1**.

Jim Leaver. 1992. *Drafting of Legal Documents*. Internal publication of the Norton Rose M5 Group of Independent Legal Practices. London, UK.

Sherman Lee. 1995. A Syntax and Semantics for Text Grammar. MPhil Dissertation, Department of Engineering, University of Cambridge, UK.

Joan Levinson. 1986. Punctuation and the Orthographic Sentence: A Linguistic Analysis. PhD Dissertation, City University of New York: New York.

E Lowe (ed). 1971. *Codices latini antiquiores: a Palaeographical Guide to Latin Manuscrits prior to the Ninth Century*. Oxford, UK.

H Löwe (ed). 1982. *Die Iren und Europa im früheren Mittelalter*. Stuttgart, Germany.

W C Mann and S Thompson. 1986. Rhetorical Structure Theory: Description and Construction of Text. University of Southern California, Information Sciences Institute, Technical Report No. RS-86-174.

Albert Markwardt. 1942. *Introduction to the English Language*. Oxford University Press: New York.

W Somerset Maugham. 1874–1965. a quotation attributed to him on URL: http://www.jumbo.com/

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

John McDermott. 1990. *Punctuation for Now*. Macmillan: London, UK.

Laurie Metcalf. 1989. *Uncle Buck*. Motion Picture: Universal City Studios Inc.

Charles Meyer. 1987. *A Linguistic Study of American Punctuation*. Peter Lang: New York.

MHRA (Modern Humanities Research Association). 1991. *MHRA Style Book — Notes for Authors, Editors and Writers of Theses: 4th Edition*. MHRA: London, UK.

R Müller. 1964. *Rhetorische and syntaktische Interpunktion, Untersuchungen zur Pausen bezeichnung im antiken Latein*. PhD Dissertation, University of Tübingen, Germany.

J Naveh. 1982. *Early History of the Alphabet*. E.J.Brill.

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

David Palmer and Marti Hearst. 1994. Adaptive Sentence Boundary Disambiguation. UC Berkeley Computer Science Technical Report, No. UCB/CSD-94-797.

Malcolm Parkes. 1992. *Pause and Effect — An Introduction to the History of Punctuation in the West*. Scolar Press: Aldershot, UK.

Eric Partridge. 1953. *You Have a Point There: A Guide to Punctuation and Its Allies*. Hamish Hamilton: London, UK.

Eric Partridge. 1954. *The Concise Usage and Abusage*. Wordsworth Reference (1995 edition: "The Wordsworth Book of Usage and Abusage"): Ware, UK.

Elsa Pascual. 1993. The Modelling of Text-Formatting for Text Generation. In *Proceedings of the 4th European Workshop on Natural Language Generation*, pages 167–171.

Elsa Pascual and Jacques Virbel. 1996. Semantic and Layout Properties of Text Punctuation. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 41–48, Santa Cruz, California, June — also available as (Jones, 1996c).

Fernando Poyatos. 1994. *La Comunicación No Verbal*. Istmo: Madrid, Spain.

Private Eye Magazine. 1995. Cartoon by Linden. *Private Eye Magazine* 16th June, page 6. Pressdram Ltd.

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

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman Group: London, UK.

P Ricci. 1943. L'interpunzione del Petrarca. *La Rinascita* **vi**, pages 258–291.

S Richardson. 1748. *Clarissa or the history of a young lady*. London, UK.

Ethel Grodzins Romm. 1984. Persuasive Writing: Comma Sense. *ABA Journal, The Lawyer's Magazine* **70**, pages 132–133. American Bar Association.

Geoffrey Sampson. 1992. Book Review of "Geoffrey Nunberg: The Linguistics of Punctuation". *Linguistics* **30**, pages 467–475. De Gruyter: New York.

Geoffrey Sampson. 1995. *English for the Computer*. Oxford University Press: Oxford, UK.

Bilge Say and Varol Akman. 1995 Information-Based Aspects of Punctuation. Unpublished Manuscript. Bilkent University, Ankara, Turkey.

Bilge Say and Varol Akman. 1996. Information-Based Aspects of Punctuation. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 49–56, Santa Cruz, California, June — also available as (Jones, 1996c).

Denise Schmandt-Besserat. 1986. The Origins of Writing. *Written Communication* **3(1)**, pages 31–45.

Denise Schmandt-Besserat. 1989. *Two Precursors of Writing: Plain and Complex Tokens* In (Senner, 1989b).

Wayne Senner. 1989a. *Theories and Myths of the Origins of Writing: A Historical Overview.* In (Senner, 1989b).

Wayne Senner (editor). 1989b. *The Origins of Writing*. University of Nebraska Press: Lincoln, Nebraska.

E Serrano. 1982. *Estudios de teoria ortografica del Español.* Murcia, Spain.

George Bernard Shaw. 1924. ...in a letter to T E Lawrence, dated 7th October. In (McDermott, 1990).

R B Sheridan. 1751–1816. ...on being asked to apologise for calling a fellow MP a liar. In (Jarvie, 1992).

Peh Li Shiuan and Christopher Ting Hian Ann. 1996. A Divide-and-Conquer Strategy for Parsing. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 57–66, Santa Cruz, California, June — also available as (Jones, 1996c).

Marthe Simard. 1993. Étude de la Distribution de la Virgule dans les Phrases de Textes Argumentatifs d'Expression Française. Master's Thesis, Université du Québec à Chicoutimi:Québec, Canada.

Marthe Simard. 1996. Considerations on Parsing a Poorly Punctuated Text in French. In *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*, pages 67–72, Santa Cruz, California, June — also available as (Jones, 1996c).

Reginald Skelton. 1949. *Modern English Punctuation: 2nd Edition*. Isaac Pitman and Sons: London, UK.

S Tannenbaum. 1931. *The Handwriting of the Renaissance*. London, UK.

L J Taylor and G Knowles. 1988. *Manual of Information to Accompany the SEC Corpus*. University of Lancaster.

R L Trask and L D Wale. 1993. *How to Punctuate*. Unpublished Manuscript, University of Sussex, Brighton, UK.

Keith Waterhouse. 1994. A Message from the AAAA's President. *The Spectator* **16 April**, page 25.

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