Exploring the Variety and Use of Punctuation

Bernard Jones
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
United Kingdom
bernie@cogsci.ed.ac.uk

Abstract

Several studies have indicated that NLP could benefit from the inclusion of a treatment of punctuation. The main impediment to the construction of any such implementation is that there no theory of punctuation upon which to base it. More basically, little is currently known about just what punctuation marks exist, how much they are used, and how they interact with each other. This study aims to answer these basic questions through the analysis of a very large corpus, and some suggestions are made for the formulation of a theory of punctuation.

Introduction

In the field of natural language processing, punctuation has been almost universally ignored, with perhaps the single exception of the period[1] marking the end of sentences. The reason for this apparent shunning is quite simple: there are no good, solid theories of the form and function of punctuation upon which to base any treatment in any computational system or linguistic theory. Therefore most current systems will simply strip out any punctuation occurring in text to be analysed or generate text that contains no punctuation marks.

Intuitively this seems very wrong. Punctuation is undeniably an integral part of written language — it would be almost unthinkable to have a newspaper article or scientific paper devoid of all punctuation other than sentence breaks, for example. Therefore it is likely that any system ignoring these extra cues within the text will suffer from reduced performance, especially if the text to be processed is *real*, i.e. the more complex sentences found in corpora from real-world sources.

Several studies have already shown the potential benefits of utilising punctuation. The study by Dale (1991) has shown the potential for punctuation in the fields of discourse structure and semantics. He suggests that punctuation marks can not only indicate degrees of rhetorical balance and aggregation between juxtaposed textual elements, but also that some punctuation marks can actually suggest the rhetorical relations that hold between such elements. Additionally Dale has shown by using punctuation in *real* sentences that discourse structure below sentence level is a reality, which prior emphasis on spoken material had deceptively missed.

[1]Throughout this paper I shall refer to sentence-final dots as periods rather than full-stops, to avoid confusion.

Our study similarly shows the potential for the use of punctuation in syntax and parsing (Jones, 1994b). A comparison was made between the performance of two grammars, identical except that one made use of punctuation and the other ignored it. For sentences that were simple in syntactic structure and punctuation use, results were similar between the two grammars. However, for the more complex sentences the punctuated grammar yielded numbers of parses that were typically two orders of magnitude smaller than those from the unpunctuated grammar. In addition to this improved performance, the study showed that use of punctuation in the grammar gave parsing results that better reflected the linguistic structure of the sentence.

In a further analysis of the results of this study, (Jones, 1994a), we showed that while there was no good relationship between the number of punctuated parses and overall sentence length or number of punctuation marks in the sentence, there was a relationship between the average length of unpunctuated lexical segments in a sentence (how many words occur between the punctuation marks in a given sentence, on average) and the number of punctuated parses of a sentence. This would seem to reinforce the arguments for making the maximum amount of use of punctuation in syntactic text analysis.

If the conclusions of these various studies are believed, then it must be a priority to generate some theory on which implementations of punctuation can be based. However, before any investigations are carried out into the syntactic and semantic functions of punctuation, i.e. the interaction of punctuation and the surrounding lexical items, it is first necessary to investigate the punctuation marks themselves. Whilst most of us could quickly produce a list of 'obvious' punctuation marks, lists from different people will not necessarily correlate. Furthermore, it is not clear which punctuation symbols are used in the texts we are likely to be analysing. An important conclusion from our earlier studies was that the ad-hoc punctuation grammar used had extremely poor coverage of punctuation phenomena. Hence it seems necessary to determine just which punctuation symbols out of the whole set are likely to be used by different categories of text producers. A University academic, for example, is likely to have a very different usage to a high-school student. This paper describes an investigation of what constitutes punctuation in real text, which symbols are used, how they interact and how frequently they occur.

Punctuation

Punctuation, as it is considered by most people, can be defined as the range of non-lexical orthography. This definition includes a very wide range of phenomena, from the sub-lexical (hyphens, apostrophes) through the inter-lexical punctuation marks to stylistic devices such as underlining and italicising, and structural devices such as paragraphing and bullet-point itemisation.

Since the super-lexical devices are still hard to represent in a computational system (since they are chiefly visual in orientation) they are not considered here. The sub-lexical marks are not terribly interesting, since their function and application is well understood and straightforward. Therefore they are also be ignored here, except where they interact with the inter-lexical punctuation marks, e.g. when they occur in word-final positions.

This investigation therefore focuses on the central portion of the range of non-lexical text-phenomena which includes the familiar punctuation symbols shown in (1). Specifically, all symbols occurring between adjacent lexical items are of interest, and therefore it is quite possible that the inter-word space could be considered as a punctuation mark. Whilst such an observation is of limited use in a language such as English, for languages like German, where compound words are often concatenated, it is quite possible that careful consideration of the space symbol could be useful.

(1) . , ; : ! ?

The literature on punctuation is not plentiful, but can be divided into two categories: the stylistic and the academic. The former category is the larger but is of limited relevance to this study. Good examples of so-called 'style guides' are those by Jarvie (1992) and Partridge (1953), but whilst their coverage of conventional punctuation is impressive, they do not mention many of the less common marks that can be found in real corpus examples. In addition, these guides tend to be rather prescriptive, ruling out many constructs that can be observed in real text.

Academic books, specifically those by Meyer (1987) and Nunberg (1990), set out to be analytical and descriptive rather than prescriptive. However, the punctuation coverage in these books is fairly limited. Meyer only considers the marks in (1) and parentheticals whilst Nunberg augments these with various quotation marks and considerations of structural devices.

Therefore it seems necessary to determine the true extent and variety of punctuation marks in realistic text, and to discover how these marks interact with each other. This information could then be used to help determine the linguistic form and function of punctuation marks within text — a prerequisite for the formulation of any theory and implementation of punctuation for NLP.

Procedure

A varied set of machine-readable corpora were processed to extract the punctuation patterns from sentences. For each sentence in a corpus, adjacent lexical items that were not separated by punctuation were replaced by a marker. The punctuation patterns, consisting of all the marks of punctuation (all non-alphanumeric inter-lexical characters) in a sentence, with lexical markers inserted appropriately, are then collected. Thus the punctuation pattern for the previous sentence would have been as in (2).

(2) e , e (e) e , e , e .

(3) e ; e , " e , e . "

Sentence closure is detected by the presence of a period (or other sentence-final marker, such as the question mark or ellipsis (...)) unless it occurs within a set of bracketing characters or quotation marks. In such a case, the end of the delimiting structure must be reached before a sentence closure can be triggered. To satisfy the principles of quote and bracket transposition (Nunberg, 1990), a closure marker immediately followed by a final delimiting mark is treated as valid, despite the apparent nesting violation (3). If the closure marker is followed by any punctuation mark other than a delimiting one or another closure marker, then it is not treated as valid. Blank lines, or end-of-files are treated as valid sentence closures and any nesting information for bracketing or quotation is reset in these circumstances.

The punctuation patterns so obtained are processed to produce the frequency of occurrence for each pattern in a given corpus, and then the patterns are broken down into their constituent symbols to produce the total frequency of each punctuation character. Note that a punctuation character is not necessarily the same as a punctuation mark, since some punctuation marks may be composed of two or more punctuation characters.

Both these sets of frequency information are reported and discussed in the results, especially contrasts and similarities between results for different corpora. In addition, the original punctuation patterns themselves are discussed with reference to conventional expectations for punctuation use.

Table 1: The Corpora used for punctuation extraction.

Corpus	Corpus size (words)	Corpus size (sentences)	Words per sentence
The Guardian '90	23,963,515	961,604	24.9
The Guardian '91	21,638,956	879,438	24.6
Leverhulme	355,594	15,547	22.8
The Bible	820,731	30,021	27.3
Philosophy (IPPE)	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	83,824,058	4,223,331	19.8

Corpora

The corpora used, with their sizes and average sentence length, are shown in Table 1. The corpora together total almost 84 million words — over four million sentences — and

Table 2: Mean distance (wds) between similar punctuation characters. (• represents numbers above 10,000)

	Guardian 1990	Guardian 1991	Lever hulme	Bible	IPPE	Proj. Gut.	Use net
.	22	22	22	31	16	22	15
?	919	790	859	249	494	244	146
!	5556	3311	1357	2622	2617	318	185
,	19	19	31	12	16	13	20
:	369	329	893	65	131	97	81
;	518	535	835	81	359	86	320
-	1724	341	256	•	47	133	38
(	364	380	267	3714	72	183	75
)	363	377	257	3714	70	177	66
[	•	•	•		978	1640	479
]	•	•	•		1110	1613	459
<	•		•		9090	•	460
>	•	4444	•		1400	5328	233
{	•		•		•	2575	2797
}	•				8933	2680	2848
‘	74	199	•		784	848	803
’	69	40	210	3616	305	349	280
"	•	•	148		491	61	74
#	•	•	•		6923	831	433
*	•	•	9118		415	622	115
^	•				8933	•	319
|	•		•		2927	•	523
/	•	•	3521		2186	5956	421
\			•		3454	•	907
_		•	4445		168	325	166
&	6015	7037	7730		1134	•	1101
%	•	•	2044		7971	165	923
$	3547	3711	•		•	1264	437
+	•	•	1539		•	1729	334
=	•	•	5229		594	510	202
@		•	•		•	•	1026
~			•		•	•	1014
tab			55		875		
Ttl	7	7	8	7	4	5	4

come from a wide range of sources. The two full years of The Guardian (a British daily newspaper), the King James Version of the Bible and texts from Project Gutenberg (an online electronic text initiative)[2] constitute the more 'formal' portions of the data. These have been produced with the help of editorial style guides/policies (so a more formal, constrained use of punctuation is expected), whilst the other three corpora have been produced more freely, without the help, or hindrance, of such guides.

Of these three freer corpora, the Philosophy corpus (a collection of philosophical papers from the IPPE initiative) is likely to be the most formal, and the Usenet corpus[3] is likely to be the least formal. The Leverhulme corpus consists of essays by secondary-school children (11–18 years old), and hence is a corpus of text that has been produced in a formal setting, but by learners. Hence these corpora should yield information regarding the use of punctuation by learners, and use of punctuation in formal and informal settings by writers who (should) know how to punctuate.

Quantitative Results

The most useful way of presenting frequency results for punctuation characters is their frequency of occurrence with respect to the lexical items surrounding them[4]. Table 2 shows the average number of lexical items occurring between each instance of a particular punctuation character. The analysis has shown 33 different printing punctuation marks. It could be argued that some of the punctuation characters in the table should be lexicalised since they (usually) have very specific meanings when used in text. This is particularly true for those shown in (4).

(4) # & % $ + = @ ~

Punctuation frequency

Punctuation frequency information is one of the most important sets of results that can be obtained from this study. It will obviously be more important to develop an interpretation for a very frequent punctuation mark than for one that is only likely to occur once in every 1000 sentences.

The comma seems to be the most popular symbol overall. In most corpora it occurs at least once per sentence, excepting the Leverhulme corpus where the probability of a comma is 0.75 per sentence and Usenet, where the probability is 0.67.

Stress markers (? !) are relatively infrequent compared with normal dots. The probability of a question-mark varies from 0.11 to 0.03 per sentence across corpora, and the variation for the exclamation mark is from 0.09 to 0.01. Although they are least frequent in the style-guided material and are most frequent in informal material such as Usenet, the frequency of these symbols seems to reflect the genre of the material more than its formality — Project Gutenberg, for example, is very rich in them. Interestingly, the Bible has the highest frequency of question marks and one of the lowest for exclamation marks, which is a reflection of its particular content rather than anything else. The question mark is more frequent than the exclamation mark in all the corpora, but the difference in frequency between the two narrows in the freer corpora.

The bracketing and quotation characters seem mostly to be equal in frequency between opening and closing characters, confirming their matching role. Minor discrepancies can be explained by missed openings or closings (either by author or analysis), but the larger discrepancies are due to use of symbols other than in their matching roles. For example, the higher percentage of closing single quotes than opening ones in many corpora is due either to the use of closing quotes to

[2]But it should be stressed that such data-rich Project Gutenberg files as π to a million decimal places have not been included!

[3]This corpus was collected from Usenet spool files from all newsgroups except those in the *comp* hierarchy. Thanks are due to Steve Finch for permission to use and modify his *grab* program to perform this extraction.

[4]The raw numerical results are uninteresting since the corpora are of wildly different sizes, and frequency per sentence is affected by varying sentence length.

mark both beginning and end of a quotation, or to an abundance of word-final apostrophes. Frequencies again tend to be rather more dependent on genre than formality of corpora. The journalistic corpora, for example, have a set of quotation marks every three sentences, on average, reflecting a high instance of reported utterances.

The characters that correspond to the remaining items of point punctuation — the colon, semi-colon and dash — occur with greater frequency in most of the corpora than the unusual symbols in the lower portion of Table 2. They also tend to be more frequent than the stress-markers. In most of the corpora the colon occurs more frequently than the semi-colon; the dash occurs with unpredictable frequency though, from only two occurrences in the entire Bible to a high of almost 10% of the punctuation in Usenet. Once more, the frequency of the symbols seems dependant on the genre of the corpora.

The only non point symbols that occur with any regularity are the various matched bracket and quotation symbols (the type used varies between corpora); the asterisk, equals sign and underscore in the Usenet, Philosophy and Gutenberg corpora; and one-offs such as the percent sign in the Gutenberg and the caret in the Usenet corpus.

Spread

The spread of symbols varies considerably between the corpora. Almost all the corpora contain all the punctuation marks shown in the top three sections of Table 2, albeit in varying frequencies. The Bible is the single exception — it contains only a few instances of the matching punctuation symbols in the third section of the table. This is presumably because of its highly standardised and edited nature.

The more unusual punctuation symbols in the lower portion of the table do not occur as standardly in all corpora. The Bible and Guardian corpora contain fewer types of these symbols than the other corpora, and the symbols the Guardian corpora do contain are far less frequent than in other corpora. The corpus that contains most of these unusual symbols is the Usenet corpus, which perhaps reflects the danger of letting largely untrained people loose on keyboards that contain too many pretty characters!

Stylistic Quirks

Some stylistic differences between corpora also emerge from the results. It is clear from the figures that, for example, the Guardian uses single quotation marks (which happen to be concatenated, forming double quotes) rather than the double-quote symbol, and that this situation is reversed in the other corpora, except the Bible, which does not contain quotation marks at all. From Table 2 it appears that at least one type of quotation character is used in the Bible, but this is an anomaly that has been caused by the presence of word-final apostrophes.

General Observations

Some corpora appear to have more than one sentence-final character per sentence. There are several explanations for this: the characters could be compounded (... !!!) or sentences could contain several sub-sentences inside delimiting structures. Stress markers can also legitimately occur within sentences, stressing particular words or phrases.

Table 3: Frequencies of increasingly complex sentences.

Commas per sentence	The Guardian 1990		The Guardian 1991		Leverhulme	
0 (e.)	64,599	44%	80,606	43%	3,940	63%
1 (e,e.)	38,882	26%	48,273	26%	1,440	23%
2 (e,e,e.)	27,282	18%	35,123	19%	628	10%
3 (e,e,e,e.)	10,759	7%	13,747	7%	188	3%
4	4,048	3%	5,274	3%	53	1%
5	1,541	1%	2,012	1%	22	
6	569		750		6	
7	234		306		4	
8	98		127			
Total	148,012		186,218		6,281	

While the comma and dot are the two most popular characters, their relative importance varies greatly. In the style-guided corpora the comma is more important than the dot, occurring from 14.4% more frequently (The Guardian, 1992) to 170% more frequently (the Bible has almost three times more commas than dots). The freer corpora produce more varied results: the Philosophy corpus is almost at parity; the Usenet corpus has 33% more dots; and Leverhulme corpus has 43% more dots than commas. Note that we are referring to dots here rather than periods, since these are punctuation symbols we are reporting, not marks. Some of these dots may be combined into other marks (e.g. ellipsis), so not all will act as periods.

Sentential Punctuation

The average number of punctuation symbols per sentence varies between corpora from 2.89 in the Bible to 4.41 in Project Gutenberg. Similarly the average number of words between (any) punctuation symbols changes across the corpora. It decreases from the formal corpora to the less formal ones, the exception being the Leverhulme corpus, which has the longest distances between punctuation.

Table 3 shows the relative frequencies of increasingly complex comma-separated multi-clausal sentences in sections of the Guardian and Leverhulme corpora, proving that increasing punctuational complexity corresponds to decreasing occurrence in the corpus. Similar results are observed with sentences containing colons and semi-colons in addition to commas.

Table 4 shows the twenty most popular sentence patterns in all the corpora. It is interesting to note the unconventional, unstopped patterns that appear, and also the isolated dots, which are probably caused both by spacing out ellipsis marks and separating the sentence-closing punctuation mark from the end of the sentence.

Table 4: The 20 most frequent sentence punctuation patterns.

Rank	Pattern	Rank	Pattern	Rank	Pattern
1	e.	8	e:e.	15	e,e,e,e,e,e.
2	e,e.	9	e!	16	e(e)e.
3	e,e,e.	10	"e,"e.	17	e,e?
4	e,e,e,e.	11	e"e"e.	18	e...
5	e?	12	.	19	e;e.
6	e	13	e:	20	"e."
7	e,e,e,e,e.	14	e(e)e:		

Table 5: Some anomalous corpus punctuation patterns.

!!:(	/=	''	!"->	::	,&	!'&)
.?!	.-	.(-:	:"	!,:	,-'	!-:
"{"	!.	-'.	!."	!...	!:"^	!:(
:'.	!:)"	!:),	=!'	,,	!:-(	:)
?:-)	,.	!;->	."	!:-|	-")	!;-)

Qualitative Results

Table 5 shows some of the anomalous punctuation concatenations that have been extracted from the corpora. Whilst some of these patterns can instantly be recognised as mistakes [,. ,,] that are mainly typographical (as in (5)), most of the other patterns in the table constitute idiosyncratic or quirky usage. Patterns that would be judged 'incorrect' by the majority of readers [.?! ." ,&] due to conflict of meaning or omission of whitespace can still have some meaning extracted from them. There is also the whole class of quirky, non-standard uses of punctuation, where particular marks used in a particular context will have some novel meaning [/= ::] that, if recognised, can be extracted. Perhaps the most typical such phenomenon observed in these results is the prevalence of the so called 'smileys', used particularly on Usenet [!;-> ?:-) .(-:]. Such marks may be interesting from a sociological perspective, but it also seems that they should be considered as punctuation, since their use is increasing in certain genres of text! Hence in analysis, we have to be careful about considering particular symbols or patterns 'incorrect', since very few of them are actually mistakes. Most symbols or patterns used will simply be idiosyncratic or unusual, so it is worth trying to extract as much information from them as is possible in the circumstances.

(5) I will begin at the most obvious,. though not necessarily the most simple level [...]

Some of the more valid, uncontroversial punctuation combinations are shown in Table 6. Here it is worth noting that whilst all the patterns are 'valid' in some sense of the word, not all are the types of patterns one would expect to find in a normal text. [@ %. # +- <<<] are examples of these unusual punctuation patterns that one might expect to find in certain genres of text (e.g. financial reports), but would be surprised to find in, say, a novel. Furthermore, even of the more conventionally linguistic patterns there are some unusual ones, that would only appear normal within specific contexts, e.g.

Table 6: Sample 'valid' corpus punctuation patterns.

!	!!!	!"-	!"?	!)	" "	"):
",	#	$)	%.	&	'	'};
(	()	('	)'?	!"'	!"?	\
))	),'	)}.	*	+	,"	,(|
,)	,_	-,	-:	."")	...	.}
/	+-	:	:-	;'	<<<	>".
?')	?;	@	[?]	["	].	_,_

[!"? [?] ,(|]. Thus there is an argument for considering all the various marks of punctuation, including those that are deemed to be genre-specific, since these will very often carry a great deal of meaning.

Sentential Punctuation

It is clear that some of the more prescriptive parts of punctuation theories are violated by examples occurring in real text. One of the few prescriptive parts of Nunberg's work, for example, states that a colon-expansion (text following a colon) cannot contain another colon but examples (6) and (7) violate this. Also the principle of quote transposition does not appear mandatory (8) (11), parentheses can be nested (9), stress-markers are concatenated (10) and nested quotation marks do not have to alternate between single and double quotes (11).

(6) Therefore the eye counters this [...] 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 [...].

(7) 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.

(8) "Sir, she fancies you".

(9) The Alchemist, by Mark Illis (Bloomsbury, 13.95 Pounds (pds)).

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

(11) Says Guthrie: "Iceblink Luck's just a good single. [...] I'm amazed people think this record's a conscious attempt to be more "coherent." It's not a conscious attempt to 'do' anything."

Conclusions

There are several main points that can be brought out of this study. The first is the general unsuitability of prescriptive theories of punctuation for text analysis. Such a wide range of punctuation occurs in real text that almost any prescriptive theory will clash. Whilst such theories are of definite use in

the area of text generation, for text analysis a far looser system is needed. When a punctuation symbol is encountered, whether it is in an unusual position or not, it should be associated with a set of possible functions or meanings that can be used to assist the linguistic analysis. Similarly unusual punctuation mark combinations should be processed to extract the maximum amount of information possible, rather than ignoring those patterns not licensed by a prescriptive theory.

What is more surprising is that the more linguistic, analytical theories, such as Nunberg's (1990), have also made the error of veering into being over-prescriptive. Hence while they are of undoubted use for guiding the development of a true punctuation theory, any prescriptive parts of these theories should be ignored as much as possible in the construction of such a theory.

The second point concerns the variety of punctuation marks we are likely to encounter. It now seems logical to divide the set of possible punctuation marks into two. This study has confirmed that there is a core subset of punctuation characters that account for the majority of instances. This subset of punctuation (the marks in (12), a parenthetical device and a quotational device[5]) could be implemented across NLP systems in a standardised manner, and cover the majority of punctuation encountered. Such a conclusion is not as obvious as it might seem, as the different sets of punctuation considered by the studies of Meyer (1987) and Nunberg (1990) show.

(12) . , ; : ! ?

The study has also shown that there exists a second subset, formed of more unusual punctuation symbols which often have a high semantic content, e.g. [% + = $]. Since the variety and particular use of these symbols is likely to be corpus specific, it is unlikely that a standardised interpretation is possible. The existance of this second set is not at all obvious, since it varies between sources and often occurs far less frequently than the main set. However, recognition of the existance of this set, and correct analysis/use will be crucial in any proper theory of punctuation. Hence punctuation treatment for any system should consist of the standard and the corpus-specific parts, which together should account for all the punctuation encountered.

Since every sentence of the English language is likely to contain 3 or 4 punctuation symbols, the argument for studying punctuation and including it in NLP is strengthened. The results further suggest that formal, edited writing produces more highly-structured material than less formal writing, in that punctuation is more frequent and more varied. The learners (Leverhulme) appear to produce text with the least complex punctuational structure. Free and learner writing also seem to include more stress markers than the formal style, suggesting either higher emotive content or symbol overuse. Hence inclusion of punctuation into analysis systems is likely to have a smaller impact if the analysed material is of a less formal nature, and least impact if learners material is to be examined. However, the punctuation that is present can still be crucial to the interpretation of the text, and so should still be considered. Since the majority of corpora consist of formal, edited material, however, we can see that use of punctuation has the capacity to be highly beneficial to analyses.

The foundation has now been laid for the development of a full theory of punctuation. Initially, instances of punctuation symbols in grammatical sentences must be studied to determine the ways in which the punctuation interacts with the lexical items surrounding it and the syntactic stucture of the sentences containing it. In this field the first set of conventional punctuation symbols is likely to produce the most useful results. In addition to syntactic investigations, similar work should be carried out to determine the semantic form and function of punctuation, and it is in this area that the second, less common set of punctuation symbols will prove very useful. From this information it should be possible to synthesise a theory of the linguistic function of punctuation marks, that can then be integrated into other linguistic systems to greatly enhance their performance, especially when dealing with real, complex sentences of natural language.

[5] Since there are several different orthographic methods of implementing these, which vary between texts.

Acknowledgements

This work was carried out under a (UK) Economic and Social Research Council studentship. Thanks for instructive and helpful comments to Henry Thompson, Alexander Holt, Andrew Fordham and anonymous reviewers.

References

Dale, R. (1991). Exploring the Role of Punctuation in the Signalling of Discourse Structure. In *Proceedings of the Workshop on Text Representation and Domain Modelling* (pp. 110–120). Technical University Berlin.

Jones, B. (1994a). Can Punctuation Help Parsing? Esprit Acquilex-II Working Paper No. 29. Cambridge, UK: Cambridge University Computer Laboratory.

Jones, B. (1994b). Exploring the Role of Punctuation in Parsing Real Text. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)* (pp. 421–425). Kyoto, Japan.

Jarvie, G. (1992). *Chambers Punctuation guide*. Edinburgh, UK: W & R Chambers Ltd.

Meyer, C.F. (1987). *A Linguistic Study of American Punctuation*. American University Studies, Series XIII. New York: Peter Lang.

Nunberg, G. (1990). *The Linguistics of Punctuation*. CSLI Lecture Notes **18**. Stanford, CA: CSLI.

Partridge, E. (1953). *You Have a Point There (A Guide to Punctuation and Its Allies)*. London, UK: Hamish Hamilton Ltd.