Towards a Syntactic Account 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

Little notice has been taken of punctuation in the field of natural language processing, chiefly due to the lack of any coherent theory on which to base implementations. Some work has been carried out concerning punctuation and parsing, but much of it seems to have been rather ad-hoc and performance-motivated. This paper describes the first step towards the construction of a theoretically-motivated account of punctuation. Parsed corpora are processed to extract punctuation patterns, which are then checked and generalised to a small set of General Punctuation Rules. Their usage is discussed, and suggestions are made for possible methods of including punctuation information in grammars.

1 Introduction

Hitherto, the field of punctuation has been almost completely ignored within Natural Language Processing, with perhaps the single exception of the sentence-final full-stop (period). The reason for this non-treatment has been the lack of any coherent theory of punctuation on which a computational treatment could be based. As a result, most contemporary systems simply strip out punctuation in input text, and do not put any marks into generated texts.

Intuitively, this seems very wrong, since punctuation is such an integral part of many written languages. If text in the real world (a newspaper, for example) were to appear without any punctuation marks, it would appear very stilted, ambiguous or infantile. Therefore it is likely that any computational system that ignores these extra textual cues will suffer a degradation in performance, or at the very least a great restriction in the class of linguistic data it is able to process.

Several studies have already shown the potential for using punctuation within NLP. Dale (1991) has shown the positive benefits of using punctuation in the fields of discourse structure and semantics, suggesting that it can be used to indicate degrees of rhetorical balance and aggregation between juxtaposed elements, and also that in certain cases a punctuation mark can determine the rhetorical relations that hold between two elements.

In the field of syntax Jones (1994) has shown, through a comparison of the performance of a grammar that uses punctuation and one which does not, that for the more complex sentences of *real* language, parsing with a punctuated grammar yields around two orders of magnitude fewer parses than parsing with an unpunctuated grammar, and that additionally the punctuated parses better reflect the linguistic structure of the sentences. Briscoe and Carroll (1995) extend this work to show the real contribution that usage of punctuation can make to the syntactic analysis of text. They also point out some fundamental problems of the approach adopted by Jones (1994).

If, based on the conclusions of these studies, we are to include punctuation in NLP systems it is necessary to have some theory upon which a treatment can be based. Thus far, the only account available is that of Nunberg (1990), which although it provides a useful basis for a theory is a little too vague to be used as the basis of any implementation. In addition, the basic implementation of Nunberg's punctuation linguistics seems untenable, certainly on a computational level, since it stipulates that punctuation phenomena should be treated on a seperate level to the lexical words in the sentence (Jones, 1994). It is also the case that Nunberg's treatment of punctuation is too prescriptive to account for, or permit, some phenomena that occur in real language (Jones, 1995).

Therefore it is necessary to develop a new theory of punctuation, that is suitable for computational implementation. Work has already been carried out on the variety of punctuation marks and their interaction (Jones, 1995), showing that whilst the set of symbols that we conventionally regard as punctuation (point punctuation, quotation and parenthetical symbols) account for the majority of punctuation in the written language (and therefore could be implemented in a standardised way), there is another set of more unusual symbols, usually with a higher semantic content, which tend to be specific to the corpus in which they occur and therefore are less suited to a standardised treatment. This study also shows that the average number of punctuation symbols to be expected in a sentence of English is four, thus reinforcing the argument for the inclusion of punctuation in language processing systems.

The next step towards the development of a theory of punctuation is the study of the interaction of punctuation and the lexical items it separates, in particular the way that punctuation will integrate into grammars and syntax. The major problem of the evaluatory studies, (Dale (1991), Jones (1994), and to a far lesser extent Briscoe & Carroll (1995)), was that their coverage and use of punctuation was rather poor, being necessarily based on human intuitions and possible idiosyncrasies. What is needed therefore is a proper investigation into the syntactic roles that punctuation symbols can play, and a formalisation of these into instructions for the inclusion of punctuation in NL grammars.

2 Data Collection

The best data sources are parsed corpora. Using these ensures a wide range of language is covered; since they are hand-parsed or checked the parse will be (nominally) correct; and since there are many parsers/editors no individual's intuitions or idiosyncrasies will dominate. The set of parsed corpora is sadly very small but still sufficient to yield useful results.

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 its immediate daughter is reported, with its other daughters abbreviated to their categories, as in (1) – (3).

(1) [NP [NP the following] :] ⟹ [NP = NP :]

(2) [S [PP In Edinburgh] , [S ...] ⟹ [S = PP , S]

(3) [NP [NP Bob] , [NP ...) ,] ⟹ [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 will include instances of serial repetition (4) and lexical 'breakthrough' in cases where phrases are not marked in the original corpus (5).

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

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

These underlying rule-patterns represent all the ways that punctuation behaves in this corpus, and are good indicators of how the punctuation marks might behave in the rest of language. In the next sections we try to generalise these rule-patterns and discuss their possible implementation.

3 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 to just 137 underlying rule-patterns for the colon, semicolon, dash, comma, full-stop.

Even some of these underlying rule-patterns, however, are questionable since their incidence is very low (maybe once in the whole corpus) or their form is so linguistically strange so as to call into doubt their correctness (possibly idiosyncratic mis-parses), as in (6).

(6) [ADVP = PP , NP]

Therefore all the patterns were 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.

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 1.

By examining all (or a representative subset) of the sentences in the original corpus that yield

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 1: Underlying colon rule-patterns

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 2: Remaining colon rule-patterns

these underlying rule-patterns, the majority of them can be eliminated. The only real underlying patterns are those in table 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. It should also be noted that 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.

- S=NP:S — inappropriate because the mother category should really be NP. Instances of this pattern in the corpus (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.

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

- NP=NP:VP — all the verb phrases for this pattern were imperative ones, which can legitimately act as sentences (8). Therefore instances of this rule application are covered by the NP=NP:S rule.

  (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 — a case of misbracketing (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.

  (9) The following were neither barred nor suspended: Stephanie Veselich Enright, [...] ; Stuart Lane Russel, [...] ; Devon Nilson Dahl, [...]

It should be noted, however, that whilst all the twelve patterns in table 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), which even cite the exceptional cases found here.

- PP=P:NP — uses the colon merely to introduce a conjunctive structure (10) — possibly one which is structurally separated from the preceding sentence fragment in, say, an itemised list and that has quite linguistically complex items.

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

- VP=V:NP & VP=V:S — 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 (11).

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

- NP=NP:NP — the only instance in the whole corpus of this pattern was a book title (12). It unlikely to be used more frequently in any other circumstances.

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

- PP=PP:PP — possibly the most productive of the excepted rules, this rule pattern provides only for a colon expansion containing a clarifying PP re-using the same preposition (13). Its use is very infrequent, though.

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

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 3: Processed underlying punctuation rule patterns

- S=PP:S — an exception since the mother category is not really a sentence (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.

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

- S=VPING:NP — a unique rule pattern whose mother is not strictly speaking a grammatical sentence (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.

  (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 3. Since some of the patterns only apply in particular, exceptional cases, the number 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 in this paper are not necessarily exhaustive for the whole language.

4 Formalism

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 (16).) Therefore (17) can be postulated as a general colon-expansion rule.

(16) The cat lay there quietly: relaxed and warm.

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

The rule generalisation for semicolons is very simple, since the semicolon only separates similar items (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.

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

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 (19) is the most appropriate.

(19) $\mathrm{T} = * .$

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 (20) simply states that a dash interpolation can contain an identical category to the phrase it follows. The second rule (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 (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.

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

(21) $\mathcal{E} = \mathcal{E} - \{$ 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 OK, so long as one of the daughter categories is identical to the mother category (22a&b). 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.

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

As an extension to these results of the analysis, it is relatively straight-forward to postulate the following simple rules (23–26), 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 still necessary if this rule is to remain in its current form.

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

- For stress-markers

  (24) $\mathcal{Z} = \mathcal{Z}\ ?$ $\mathcal{Z} : *$
  (25) $\mathcal{Y} = \mathcal{Y}\ !$ $\mathcal{Y} : *$
  (26) $\mathcal{W} = \mathcal{W} \ldots$ $\mathcal{W} : *$

5 Implementation Methodology

The issue now arises of the best way to integrate punctuation into a NL grammar. There are three existing hypotheses to choose from. The theory of Nunberg (1990) is that punctuation should be treated in a 'text grammar' on a separate level to the lexical grammar. However, as pointed out by Jones (1994), it is difficult to see how this would be feasible in practice and there is little linguistic or psychological motivation for such a separation of lexical text and punctuation.

Therefore Jones (1994) fully integrates punctuation and lexical grammar, and in effect treats punctuation marks as clitics on words, introducing additional features into normal syntactic rules (27). Briscoe and Carroll (1995), however, point out that this makes it hard to extract an independant text grammar or introduce modular semantics. Therefore their grammar keeps the punctuation and part-of-speech rules separate, but still allows them to be applied in an interleaved manner, in effect finding the happy medium between the two extreme approaches. Hence, additionally, their rules include the punctuation marks as distinct entities, rather than cliticising them, although they still require extra features to ensure proper application of the rules (28).

(27) np[st $\mathcal{S}$] → np[st c] np[st $\mathcal{S}$][1]

(28) V2[WH -,INV -] →
H2[WH -,FIN +,-ta] +pco[2] V1[VFORM ING]

The most appropriate method would seem to be a combination of the two integrated methods above, combining their modularity, flexibility and power. Thus the Generalised Punctuation Rules obtained above could be encoded into a normal syntactic grammar to add punctuation capabilities. However, this will almost certainly result in overgeneration of parses, as the rules are still too flexible: they accurately describe syntactic situations where punctuation can occur, but fail to place any constraints upon those situations. Hence some further theoretical work seems to be required to constrain the applicability of these rules.

The main location for punctuation marks is likely to be with *phrasal-level* items, whether the marks occur before a particular phrasal item or after it. Punctuation does not seem to occur at levels below the phrasal, with one exception: punctuation is allowed to occur at any level in the context of coordination. Thus (29) represents

[1] $\mathcal{S}$ represents a variable

[2] +pco represents a comma

legal use of punctuation adjoining a phrasal item since it occurs adjacent to the ADJP within the NP. However, in (30) there is no phrasal item for the punctuation to attach to, and so its use is unsanctioned. Conjunctive punctuation use can be seen in (31), where although occurring below the level of NP, the punctuation is legal because of its conjunctive context.

(29) The green, more turquoise actually, bicycle ...

(30) * The, bicycle is a joy to ride.

(31) The shark, whale and dolphin can all swim.

To generalise, then, punctuation seems to have adjunctive and conjunctive functions, and the theoretical formalisation of these function will form a good method of constraining the parses produced with the Generalised Rules above.

6 Conclusion

We have seen that by extracting punctuation patterns from a corpus it has been possible to postulate a small number of generalisations for punctuation rules within NL grammars. A suitable methodology for applying punctuation to existing grammars has also been suggested. Since many of the rule patterns seem to have a very low frequency of occurrence it may also be useful to collect such frequencies and use them in the rule generalisations to attach probabilities to various rule expansions. We have also seen that the rule patterns we extracted from the corpora agreed to a large extent with the descriptions of punctuation use found in publishers' style-guides, suggesting that reference to these may be useful.

What is needed now is a thorough testing and evaluation of the suggestions made in this paper, both against punctuation patterns from other corpora and in parsing novel material, to maybe suggest better generalisations. Then the next step towards a theory of punctuation can be carried out, namely the analysis of punctuation for its semantic function and content.

Acknowledgements

This work was carried out under Research Award R00429334171 from the (UK) Economic and Social Research Council.

Thanks for instructive and helpful comments and suggestions to Alexander Holt, Henry Thompson, Ted Briscoe and anonymous reviewers.

My regards to the international academic and research community in the field of Computational Linguistics: thank-you, and good-bye!

References

Edward Briscoe. 1994. Parsing (with) Punctuation, and Shallow Syntactic Constraints on Part-of-Speech Sequences. RXRC Grenoble Laboratory, Technical Report.

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

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.

Gordon Jarvie. 1992. Chambers Punctuation Guide. W & R Chambers Ltd., Edinburgh, UK.

Bernard Jones. 1994. 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. 1995. Exploring the Variety and Use of Punctuation. In *Proceedings of the 17th Annual Cognitive Science Conference*, pages 619–624, Pittsburgh, Pennsylvania, July.

Geoffrey Nunberg. 1990. The Linguistics of Punctuation. CSLI Lecture Notes **18**, Stanford, California.